Extracting and Linking Locations and Activities from the Geospatial Web

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

Von

Ramya Venkateswaran

aus Republik Indien

Promotionskomitee

Prof. Dr. Robert Weibel (Vorsitz)
Prof. Dr. Ross Purves
Prof. Dr. Dirk Burghardt

Zürich 2015

Summary

The Geospatial Web contains large amounts of data that are waiting to be turned into information. This information is of great value to map services since it enables them to push dynamic data on to mobile devices where the location of the user is known. With location information constantly updating, this data can be used to inform the user of the actions that people perform in that place, or where most people go for various activities such as eating, watching a movie, etc.. This information could be extracted from unstructured text such as web pages in HTML or structured content on the web. However, people's perception of a place is never uniform. Their view of the extent and characteristics of a place often depend on their own experiences. The nature of a place often undergoes constant change due to new infrastructure driven by increasing population, changes in economy, seasons, etc. It would be very useful if map applications could capture these changes too.

The principal contribution of this thesis is towards establishing the link between people's perception of place and the actions they are likely to perform in that place. In this thesis we measure actions by the activities people are likely to perform. Since this research focuses on people's perception of place, the thesis investigates and demonstrates methods to semi-automatically determine activities people perform from UGC and automatically link them to locations, also extracted from UGC. Since several conclusions in the course of this work are based on UGC and the Geospatial Web, the thesis first discusses an approach towards computing the geographic and linguistic web coverage of the relevant data. It also discusses a study to examine the geographic and linguistic variations in these data. This is especially important since these data are known to be geographically and linguistically heterogeneous.

This thesis investigates various questions, some of which are: How does the geographic and linguistic web coverage vary across Switzerland for tourism related themes and how is it affected by different factors such as population and touristic popularity? How can UGC be utilised to investigate people's perceptions of places thereby extracting locations and perceived activities automatically? How can UGC be used to make affordances on a certain place? Once extracted, how do these activities relate to topographic data and how do they behave spatially and temporally?

Zusammenfassung

Das Geoweb enthält riesige Datenmengen, die nur darauf warten, in Informationen umgewandelt zu werden. Diese Informationen sind deshalb so wertvoll für Kartendienste, da sie es erlauben, dynamisch Daten auf Mobilgeräte zu übertragen, wenn die Position des Nutzers bekannt ist. Da Geodaten ständig aktualisiert werden, kann man den Nutzer darüber informieren, was andere Nutzer am selben Ort tun, oder welche Orte sie für bestimmte Aktivitäten aufsuchen; wo sie beispielsweise essen oder ins Kino gehen. Diese Informationen können entweder aus unstrukturiertem Text wie HTML Webseiten oder aus strukturierten Netzinhalten extrahiert werden. Allerdings ist die Wahrnehmung eines Ortes durch mehrere Menschen stets unterschiedlich. Der Eindruck über die Größe eines Ortes oder dessen Eigenschaften hängen stark von den eigenen Erfahrung der Nutzer ab. Auch verändert sich ein Ort durch neue Infrastrukturen, die einer steigenden Bevölkerungsdichte Rechnung tragen, durch veränderte wirtschaftliche Umstände, je nach Jahreszeit und weitere Faktoren. Anwendungen müssen daher auch diese Veränderungen berücksichtigen.

Der fundamentale Beitrag dieser Arbeit ist, eine Verbindung zwischen der Wahrnehmung eines Ortes durch eine Person und dem, was sie dort wahrscheinlich tun wird, zu schaffen. In dieser Dissertation schließen wir von den Aktivitäten einer Person auf ihre Handlungen. Da sich diese Forschung auf die Wahrnehmung eines Ortes durch die Menschen konzentriert, untersucht und entwickelt die vorliegende Arbeit Methoden, die halbautomatisch die Aktivitäten aus nutzergenerierten Inhalten (UGC) extrahiert und sie dann automatisch mit Standorten verbindet, die ebenfalls aus nutzergenerierten Inhalten entnommen werden. Der Fokus dieser Arbeit richtet sich auf die nutzergenerierten Daten und das Geoweb. Dies erfordertin einem ersten Schritt Erfassung der geographischen und sprachlich-räumlichen Abdeckung dieser Daten um eine Vorstellung zu erhalten,, welche und wieviele Daten im Geoweb zu Verfügung stehen. Die Abdeckung der Daten im Netz ist nicht homogen; sie variiert geographisch und sprachlich. Dies erfordert daher erst eine Untersuchung dieser Variationen, bevor diese Daten für weitere Untersuchungen verwenden können.

In dieser Dissertation untersuchen wir verschiedenen Fragestellungen, unter anderem: Wie verändert sich die sprachliche und geographische Netzabdeckung von Tourismusthemen an verschiedenen Orten der Schweiz und wie wird sie durch verschiedene Faktoren wie Bevölkerungsdichte und Beliebtheit bei Touristen beeinflusst? Wie kann man nutzergenerierte

Inhalte nutzen, um auf die Wahrnehmung eines Ortes durch den Nutzer zu schließen, um dann wiederum Geodaten und Aktivitäten automatisch daraus abzuleiten? Wie können nutzergenerierte Inhalte verwendet werden, um zu erfassen wie ein Ort genutzt wird und welche Aktivitäten er ermöglicht? Wie verhalten sie diese abgeleiteten Aktivitäten in Bezug auf topografische Daten und wie verändern sie sich in zeitlich und räumlich?

Acknowledgements

This work has been carried out at the Department of Geography, University of Zurich. I have a number of individuals from University of Zurich and outside to thank for helping me through this research and finally the thesis.

Robert Weibel – My first advisor, I am extremely thankful and grateful to him for helping me finally accomplish my goal. Not only his help and support as an advisor is acknowledged, but also his patience and words of encouragement helped me lead to my goal and also supported me through the not so positive times during the PhD.

Ross Purves – My second advisor whose brilliance and extremely directed comments and suggestions, helped shaping my chaotic research into a thesis into. His guidance and patience has been instrumental in leading me to my final goal.

Dirk Burghardt (TU Dresden) – Who was my second advisor during the first year, was a part of many brainstorming sessions that helped me form my research questions and continued later on with guiding me and keeping me on track.

Lorenz Hurni (ETH Zurich) – I am extremely thankful to him for agreeing to be the external reviewer for my PhD thesis.

Pia Bereuter – Whose company as an office mate and project partner made work more interesting. Her help with identifying toponym ambiguities in different languages is duly acknowledged.

Ronald Schmidt – The ArcGIS *guru* of our unit, he helped with designing and making the maps in ArcMap that are presented in this thesis.

My past and present colleagues and friends at GUIZ (Meysam Aliakbarian, Pia Bereuter, Curdin Derungs, Christian Gschwend, Peter Jeszenszky, Izabela Karsznia, Benjamin Rohrbach, Ronald Schmidt, Ali Soleymani, Georgios Technitis, Martin Tomko, Flurina Wartmann, Azam Bahrehdar, Ralph Straumann, Somayeh Dodge, Patrick Lüscher, Patrick Laube, Alistar Edwardes, Kenan Bektas, Arzu Çöltekin, Sara Irina Fabrikant, Tumasch Reichenbacher, Paul Crease, Stefano De Sabbata, Marco Salvini, Jan Wilkening and Anna-Katharina Lautenschütz) for their support, company and stimulating *znüni, zvieri* and lunch conversations.

Annica Mandola and Elisabeth Cottier, for helping me through all the administrative work and for their company.

My mother and father - I am thankful for their efforts in raising and educating me and giving me the best.

Lastly, but not the least, my two precious gems, my world – Suman and Kanika for making sure life is always happy and exciting, sometimes a bit too much!

Ramya Venkateswaran, Zurich, January 24, 2015

This research was funded by the Swiss National science Foundation (SNF) twice, for the projects GenW2 (grant no. 200020-120256) and GenW2+ (grant no. 200020-138109). Their support for both instances has been instrumental in the making of the research and this thesis, which is gratefully acknowledged.

Table of Contents

Summary ii

Zusammenfassung iii

Acknowledgements v

Chapter 1. Introduction 1

1.1 Overview 1

1.2 Motivation 2

1.2.1 Adaptive generalisation through flavour based data integration 2

1.2.2 User perspectives and affordances from place 3

1.3 Research questions 5

1.4 Structure of the Thesis 5

Chapter 2. Background 7

2.1 Introduction 7

2.2 The Geospatial Web 7

2.2.1 Web 2.0 and the Geospatial Web 8

2.2.2 Geospatial data 11

2.2.3 Geographic information retrieval 12

2.3 Exploiting and extracting information from User-Generated Content 16

2.3.1 Volunteered Geographic Information versus User-Generated Content 16

2.3.2 Knowledge extraction from image metadata 19

2.3.3 Deriving regions from the web 20

2.4 Geographic place and space, activities and affordances 21

2.4.1 Geographic place and space 21

2.4.2 Affordances and activities 22

Chapter 3. Overall Research Methodology 25

3.1 The coverage of the Geospatial web 25

3.2 Determining and describing locations and related activities from User Generated Content (UGC)28

Chapter 4. Of Web counts: Geographic coverage and linguistic differences 33

4.1 Introduction 33

4.2 Data and Methods 35
4.2.1 Approach 35
4.2.2 Toponym data 36
4.2.3 Toponym ambiguities 41
4.3 Geographic web coverage 43
4.4 Linguistic coverage 48
4.5 Influencing factors 52
Chapter 5. Determining locations and related activities using UGC 57
5.1 Introduction 57
5.2 Data 58
5.2.1 Flickr 58
5.2.2 GeoNames 59
5.3 Initial steps and activity terms 60
5.4 Methods and experiments 62
5.4.1 Automatic shrinking bounding box (ASBB) approach 63
5.4.2 Determining locations from the ASBB method 64
5.4.3 Activities and their locations 68
5.4.4 Grouping activity locations 75
Chapter 6. Taking place descriptives further: Exploration of activities 79
6.1 Introduction 79
6.2 Infrastructure v/s activity points 80
6.3 Activity interaction through co-occurrence 82
6.4 Activity behaviour using activity theory 84
6.5 Activity clusters 87
Chapter 7. Discussion 93
7.1 RQ 1 – Web coverage across Switzerland 93
7.1.1 Main contribution 93
7.1.2 Variation of Geographic web coverage 94
7.1.3 Language differences in the web coverage 95
7.1.4 Factors affecting the web coverage 96
7.2 RQ 2 – UGC and people's perceptions of places 98

7.2.1 Main contribution 98
7.2.2 Locations and activity terms 98
7.2.3 Location similarity based on activities 100
7.3 RQ 3–Activities in space and time 103
7.3.1 Main contribution 103
7.3.2 Activities w.r.t. space and topographic data 104
7.3.3 Activity interaction and behaviour 106
Chapter 8. Conclusion 109
8.1 Achievements and insights 109
8.2 Open issues 112
8.3 Outlook and Future work 113
Bibliography 115
Complete publication list 134
Curriculum Vitae 135

1. Introduction

1.1 Overview

Digital maps commonly offer users an allocentric view of the environment around them. The user then zooms in to their place of interest or current position. This interaction is further streamlined on devices such as smartphones, where the user's geographical location can be automatically derived. By means of a built-in GPS units, mobile devices are able to provide personalised, geographically-oriented data and information services across wireless networks (Shiode et al., 2004; Jiang and Yao, 2006). Furthermore, through map generalisation the user is then presented with only that information which is most relevant to them. Map generalisation is a technique used in cartographic scale reduction, where complexity is reduced by emphasising essential information and suppressing unimportant information, while maintaining logical and unambiguous relations between map objects (Weibel, 1997). Even though smartphones are considered personal devices, real-time map generalisation is often performed using similar criteria for all users, without fully making use of information such as the user's profile, habits, and location. The main motivation behind this thesis is towards amending this and making the process of real-time generalisation more dynamic and egocentric (Meng, 2005).

In order to present a user centric map or a map that is adapted to the user and their context; the user's perspective, such as their perspective of a place, its extent, location etc., play an important role. Although location based services (LBS) (Poslad et al., 2001; Zipf, 2002) may be an answer to this, this proposal goes a step further than location based services, in that it involves pre-processing in the generalisation stage, also known as background processing. Since mobile devices are often constrained in terms of network bandwidth and processing power, this proposal relies on performing data integration to inform the generalisation operators, thereby reducing the amount of computation required on mobile devices. To demonstrate the methods discussed in this thesis, tourism has been selected as a use case.

In order to represent user perspectives on a map, the need to link the concepts of space and place is necessary, for which Edwardes's (2007) space-place continuum is used in later chapters as a reference. The more the information on the user and context, the more decisions the generalisation operators are able to make, thereby influencing the generalisation process. People's perception of

a place is never uniform. Their view of its extent and characteristics often depends on their own experience (Cresswell, 2009). The nature of a place is often under constant change owing to new infrastructure driven by increasing population, changes in economy, seasons, etc. The process of generalisation often uses a spatial database known as multi-representation database (MRDB) to store data (Weibel and Dutton, 1999). Although reliable and efficient, MRDBs are often static thereby not always be kept up-to-date with such changes, and may therefore not be able to serve users with up-to-date information. In the latter part of the chapter, an approach to infer such changes is discussed in detail.

This research is driven by the vision of a convenient and time-efficient digital tourist map rendered on one's mobile device, pointing out places one might be interested in along with suggestions on what activities may be performed there. Based on user preferences, interests, and past patterns available from the user's profile, places are shown in the corresponding order of importance. The remainder of this thesis works towards this vision, but the contribution of this thesis is mainly in the form of techniques and methods that help build individual components rather than a fully functional application.

1.2 Motivation

1.2.1 Adaptive generalisation through flavour based data integration

A typical multiscale digital map consists of an MRDB that contains several geographic datasets at various resolutions, for a single area. An MRDB is a spatial database that is used to store data at different levels of resolutions, thereby serving as a pre-generalised and pre-harmonised data store for different scales (Weibel and Dutton, 1999; Hampe and Sester, 2002). While the nature of this information is often static, various algorithms are used to select the information relevant to a given map. One of the objectives of this project is to use time-sensitive information gathered from the web, a task made easier in the era of Web 2.0 (O'Reilly, 2005). The project proposes to do this through the approach of 'flavour based data integration' (Venkateswaran, 2009), a mechanism for integration of data from heterogeneous Web 2.0 sources. In this case a 'flavour' refers to a targeted application, such as tourism, sports, or culture. For recording information collected through data integration, we propose a structure that we call the FactsDB (read as facts database). The FactsDB is intended as an auxiliary dataset, which in turn contains information items (called "facts") about different places. The auxiliary data in the FactsDB can be linked to a spatial database (such as a

Multiple Representation Database, or MRDB) via a gazetteer. In this way, the spatial data of the MRDB can be enriched with additional information that, at a later stage, may inform the generalisation process by providing additional semantics about particular places. The simplicity of the FactsDB makes it amenable to frequent updates. Through this mechanism, generalisation in web and mobile portrayal can be made adaptive to the particular needs of specific user groups or 'flavours' (e.g. tourists). The research in this project mainly focuses on the nature of semantic information that can be extracted for enriching the MRDB, as opposed to the specifics of the structure of the FactsDB itself. Approaches towards data enrichment in literature on this subject have typically focused on structures and object relationships inferred from topographic data, which are commonly created by either official or commercial data producers. It has been described as the process of extracting implicitly contained information in the support of map generalisation (Neun et al., 2004).

The information in the proposed FactsDB will be extracted from the web. Effectively, we will perform data enrichment through the extraction of georeferenced data from user generated content (UGC). However, as every geographer knows, it is very unlikely that web content is evenly distributed in space. In order to generate geographic data from the web that can be used as reliable information, knowledge of geographic web coverage is important in order to establish how well certain areas are geographically covered by the web. Frequency and patterns that emerge from this data collection, aid in the decision of preselecting web data for further investigation, based on the nature of the study. Studies which seek to draw conclusions based on, for example, variations in density, must first estimate the underlying density of the collection of interest. Implicit assumptions about homogeneity of coverage can be misleading. Therefore, even before performing data integration, a study of the coverage is important and has been addressed in this thesis.

1.2.2 User perspectives and affordances from place

Mobile devices impose several constraints, such as screen size and limited computing power. Research in the latter area is progressing and the mobile hardware industry is very quickly moving towards boosting compute performance and power efficiency. Furthermore, various processor manufacturers are able to limit the size of their processors while still matching the functionality and performance of a normal desktop computer. However, screen size continues to impose a significant restriction, especially in the area of map presentation, given that digital maps on mobile devices are often cluttered and offer too much information. When users are presented with a map, there is a need for special generalisation algorithms in graphical information delivery and portrayal

that are able to reduce the information shown, while adapting to context and user profile. Mobile devices are very often personal devices; they tend to be owned and used by only one person. On such devices, personal profiles tend to be available, and when used, can inform the generalisation process on the user's habits and what the user might want to see on a map. This has not been exploited enough, and the combination of user profile and location can help filter a significant amount of information on the map and make the map contents more relevant to the user.

This thesis aims at solving a part of the challenge discussed above. As discussed earlier, UGC contains an abundance of unexplored data. Our intention is to use these data towards extracting information that can be used to describe the user's perspective and what they might want to see on a map instead of general map that might contain too much information. Furthermore, when it comes to perspective, individuals do not think about places in the same way that cartographers represent them on maps. In other words, a place is not thought of in terms of its coordinates or its administrative boundaries. Instead, individuals may have mental maps of a place, its extent, or even feelings that they associate with a place; i.e. the sense of place (Cresswell, 2009). Jordan et al. (1998) posit that the *actions* people perform are one of the aspects of place, and this is particularly true in the context of tourism, which is the use case explored in this thesis. In the past, Purves et al. (2011) have described place with the help of geographic elements (e.g. cities, rivers), qualities (e.g. old, blue) and activities (music, rafting). Speaking of actions, places afford different actions for humans. Affordances is an important concept for this thesis. The general term and theory of affordances was introduced by Gibson (1979) which are "qualities of an object or environment that communicate opportunities to do certain things". Later in this thesis, this concept is extended to places.

The activities that individuals perform in places can be extracted from UGC using various methods. In the later chapters, this thesis analyses in greater detail, the general perception of what qualifies as an activity. Another objective of this thesis is to link the behaviour and relationship indicators of various activities extracted from UGC to different generalisation operators for adaptive generalisation. Using this technique, different map products can be generated based on, for example, whether a particular region appears to be more frequently associated with cycling or hiking. Each map product could focus on elements important to these activities. The target activity dictates which information items are important, and therefore displayed and highlighted, and which items are unimportant, and can therefore be omitted.

1.3 Research questions

Against the motivation and objectives outlined above, this thesis addresses three sets of research questions which will be dealt with in chapters 4 to 6:

1. What is the web coverage across Switzerland for tourism related themes and how is it affected by different factors?

    1.1. How does the geographic distribution of web coverage for tourism-related themes vary across Switzerland?

    1.2. Are there any differences in web coverage distribution for different languages and gazetteer datasets?

    1.3. How do typical factors such as population and touristic popularity of a place affect web coverage?

2. How can UGC be utilised to investigate people's perceptions of places? How can UGC be used to make affordances on a certain place?

    2.1. With the help of UGC, can we extract locations of places and its activities? Is it possible to assign individual locations to these activities?

    2.2. Having extracted locations and their activities, is it possible to group these locations based on how similar the activities performed there are?

3. Once extracted, how do these activities behave spatially?

    3.1. How do these tourism-related activities relate to space and how do they relate to topographic data?

    3.2. How can interaction between activities be measured, what is the nature of their interaction and how do they behave spatially and temporally?

1.4 Structure of the Thesis

This thesis is organised as follows:

Chapter 2 goes into the details of the literature and provides some background about technology that is related to the research carried out in this thesis. For this thesis, topics such as the geospatial web, user-generated content, volunteered geographic information, concepts of space and place, affordances, and activities are discussed.

Chapter 3 outlines the methods according to which this research was carried out and points out the gaps in research.

The next three chapters, **4, 5 and 6** refer to research questions one, two, and three respectively and contribute the main substance of this thesis. Each research question involved a set of experiments and some analysis. **Chapter 4** has been published in TGIS as a separate journal paper (Venkateswaran et al., 2013).

Chapter 7 discusses this thesis at an aggregate level. It looks at the three research questions and its sub-questions and answers each one individually.

Chapter 8 concludes this thesis and summarizes the main achievements and insights gained in this thesis. It discusses possible directions for future work

Background

1.5 Introduction

This chapter attempts to explain the state of the art and the general area where this thesis intends to make its contributions. This thesis borrows and builds on ideas from the Geospatial Web. It is hard to estimate how old the concept of the Geospatial Web is. Wikipedia reports that the concept could have been introduced as early as 1994.

The importance and size of the Geospatial Web has grown with time and has been closely associated with Web 2.0. Then next section briefly describes both these concepts together. The sections later deal with the nature of geographic data on the web and some concepts related to Geographic Information Retrieval (GIR), which is relevant to the discussion around the extraction of information from the Geospatial Web.

1.6 The Geospatial Web

Although there is no commonly agreed definition for the Geospatial Web among researchers, many give us an idea of what the Geospatial Web could possibly include. Lake and Farley (2007) explain that "it refers to the global collection of general services and data that support the use of geographic data in a range of domain applications". Purves (2011) writes that the Geospatial Web is "any part of the Web that is somehow related to geography" and Haklay et al. (2008) state that it is the "merging of geographic information with abstract information". The terms 'GeoWeb', 'Geographic World Wide Web' and 'GeoWeb 2.0' (Maguire, 2007) have often been closely linked with the Geospatial Web and very often are used interchangeably by researchers. The web is an important and vast repository of geographical information. Hill (2006) estimates that up to 70% of text documents contain place name references, while Sanderson and Kohler (2004) suggest that 13-15% of all search engine queries contained a geographical term such as a place name, a post code, a type of place or a directional qualifier, such as "north". The Geospatial Web became widely popular when products like digital maps (Google Maps, Yahoo! Maps and Bing Maps) and virtual globes (Google Earth, NASA World Wind) became easily accessible to the general population of internet users. In 1994, The Xerox Palo Alto Research Center (PARC) publicly launched a web-based product called Map Viewer. It was an interactive world map that could create and display

maps of any part of the world on demand (Putz, 1994). Around 2005 Google, Yahoo! and Microsoft have also released free APIs which allowed users to map their own data, thereby contributing to the amount of information on the Geospatial Web. These services not only presented a map interface and map data but also supported spatial search and location based services.

On the hardware side, cheaper handheld devices and mobile phones with positioning capabilities flooded the market. Their omnipresence enabled mobile software applications to obtain real-time location information of their users. Through map based applications and products, users were able to make use of Location-based services (LBS). These services were originally introduced in the early 2000s and later became widespread simultaneously with dramatic advances in mobile computing and positioning technology. They aimed to provide personalised, geographically-oriented data and information services to mobile users based on their current location (Shiode et al., 2004; Jiang and Yao, 2006). Additionally, the changing context and dynamic location information required the mobile system to adapt visually, which led to a new branch of mobile cartography for creating adaptive maps (Meng et al., 2005).

1.6.1 Web 2.0 and the Geospatial Web

The term Web 2.0, although not very different from the World Wide Web (Berners-Lee and Cailliau, 1990), was used in the O'Reilly Media Web 2.0 conference 2004 (O'Reilly, 2005), to refer to a set of changes in the way in which developers and users "facilitate interactive information sharing, interoperability, user-centered design, and collaboration on the World Wide Web" (Berners-Lee and Cailliau, 1990, Wikipedia, 2010). One of the main advantages of Web 2.0 was its increasingly collaborative nature, which in turn created a platform of sharable information that grew over the past few years. Adding location to this information, be it in text, photos, or videos, became very popular. The plethora of georeferenced information led to the invention and widespread use of various web technologies and concepts around geospatial data integration. Increasing positional accuracy and falling costs of GPSs led to their ubiquitous presence in cell phones, handhelds, cameras, and sports equipment, contributing to the amount of georeferenced information on the web.

The predecessor of Web 2.0 was driven by complex architectures, leaving the end user with little or no control over data. Eventually, these complex server side solutions were replaced by mostly free and easy-to-use web services. Web services were a new method of communication between

devices, and promised standards-based information interoperability (Zhao et al., 2007). A web service is a modular application that is self-described, self-contained and is discoverable and accessible across the Web (Di et al., 2005).

To exchange messages between devices on the web, the Extensible Markup Language (XML) is the most commonly used language. It is a markup language whose structure facilitates ease of use and human and computer readability. It is a subset of the Standard Generalized Markup Language (SGML) and was designed by W3C (XML Specification 1.0) in an effort to create open standards (Bray et al., 1997). XML documents or files are exchanged using protocols such as Simple Object Access Protocol (SOAP) (SOAP Version 1.2) or XML remote procedure call (XML-RPC) (Laurent et al., 2001). In order to describe a web service, the Web Services Description Language (WSDL) is used. The definitions are written in XML. To enable these web services to be discoverable, the Universal Description, Discovery and Integration (UDDI) (UDDI Version 2 Specification), a directory service used by businesses to register and search for web services was introduced. UDDI uses WSDL to describe interfaces on the web and communicates using SOAP. Together, these services are known as web application programming interfaces (APIs) (Curbera et al., 2002). Web services use an XML message centric approach and very often, client side logic is written in a light scripting language, such as JavaScript. GIS-based applications tend to use a technique called AJAX (Asynchronous JavaScript and XML), which allows web-based applications to asynchronously and incrementally fetch data from a server, enabling more efficient transfer of data between the application and the server, as well as a more responsive user interface. (Sayar et al., 2006). The interactive experience is possible because AJAX makes use of the XMLHttpRequest as a messaging protocol, allowing remote information to be continuously available to the user, as opposed to only after a page refresh. Once the web page is rendered, heavy weight components do not have to be rendered again and again, and data can be transferred asynchronously. Google Maps makes use of this technology[1].

Analogous to Web 2.0, there are ongoing efforts in making the Geospatial Web ubiquitous. The Open Geospatial Consortium (OGC), a similar body to the World Wide Web Consortium (W3C), was founded in 1994. It aimed at standardising and creating an open platform for data on the geospatial web in order to facilitate better interoperability of services, documents, and

[1] http://www.adaptivepath.com/ideas/ajax-new-approach-web-applications/

heterogeneous data. The OCG proposed an architecture framework known as the OpenGIS Reference Model (ORM, 2011), which describes a framework of their specifications for implementing interoperable solutions and applications for geospatial services and data. To share geographic information across applications, both web and desktop, OGC defined an XML-based language known as Geography Markup Language (GML). Extensibility is intrinsic, as GML is written in XML schema and the language contains a collection of primitives for encoding time, geometry, topology, coordinate reference systems, units of measure, map styling, observations coverages, and other geographical features (Lake and Farley, 2007). Along similar lines, Google popularised Keyhole Markup Language (KML), which is also an XML based language, to encode geographic features. Although KML is very similar to GML it was tailored for use with Google Earth and is better adapted for dynamic features and loading content over the Web (Honda et al., 2006). Finally in 2008[2] it was adopted as an OGC implementation standard and has since been widely used in web and desktop related GIS applications.

GIS-related web services are often available to users in the form of simple APIs. With these APIs, web users are often able to represent georeferenced information from many sources onto a map, with the help of simple map APIs provided by Google, Yahoo and Microsoft, thereby creating *Mashups*. Map mashups (Butler, 2006) are not only useful to web users but also to researchers in the academia. This technology soon led to a small step in spatial decision making (Rinner et al., 2008). Tran (2007) in his blog titled "Google Maps Mashups 2.0" estimated that after the official release of the Google Maps API in June 2005, web users created over 50,000 Google Maps mashups in a period of approximately two years. This combination of information with the application of basic geographic techniques and mapping, led Turner (2006) to coin a new term: neogeography, also known as new geography. Turner (2006) states that "Neogeography is about people using and creating their own maps, on their own terms and by combining elements of an existing toolset. Neogeography is about sharing location information with friends and visitors, helping shape context, and conveying understanding through knowledge of place". Map making, which initially was a forte of, and limited to, geographers and cartographers, very quickly became something that neogeographers participate in, producing maps in a matter of minutes. This led to a boom in the availability of Mashups on the web. Neogeographers are "those with little or no formal geographic training who contribute geographic information on a voluntary basis using the technologies loosely known as Web 2.0" (Sui and DeLyser, 2012). However, there is considerable

[2] KML - http://www.opengeospatial.org/standards/kml

debate in the geographic community on the information they produce and their "disregard to existing geographical and cartographical traditions" and/or their "lack of understanding of spatial analysis" (Haklay et al., 2008). Geographers have not been convinced of the practices and visualisation methods used by neogeographers, but they all agree that the information generated by them is of considerable use to the community. Neogeography is closely related to the concepts of User Generated Geographic Content and Volunteered Geographic Information, which are discussed in the Section 1.7.1.

1.6.2 Geospatial data

The amount of geospatial data on the web is large. Typically, any web content with a reference to a geographic object is a candidate to be part of this large geospatial repository. Mark et al. (1999) define a geographic object as a spatial object on or near the surface of the earth, some examples of which include regions, land, water bodies, mountains and canyons, hills and valleys, roads, buildings, and bridges. Geospatial data is anything that identifies or refers to these objects. An online digital gazetteer or web pages describing places to visit are examples of geospatial data on the web.

Digital gazetteers have been defined as "geospatial dictionaries of geographic names with core components of a name, a location and a type" (Hill, 2000). Some examples of online gazetteers are the gazetteer server developed by Alexandria Digital Library (ADL), The Getty Thesaurus of Geographic names, GeoNames, and the U.S. Geological Survey's (USGS) Geographic Names Information System (GNIS). A gazetteer is an example of structured geographic information. Although this information is structured, there is not much standardisation around the structure of a gazetteer, which in turn made interoperability between gazetteers difficult. A number of researchers have discussed interoperability (Kessler et al., 2009).

In addition to textual and gazetteer data, georeferenced images such as those in Flickr, Geograph, and Picasa, georeferenced videos such as those on YouTube, online maps such as Google maps and Bing maps also qualify as geographic information and are also part of the Geospatial Web. However, in order for Geographic Information Systems (GIS) to make sense of this data and render it in the form of geographic objects, it is essential for this data to be more structured. Retrieving information and making sense of this data in order to make it georeferenced and identifiable in the form of coordinates and topology, are dealt with by methods that fall under Geographic Information Retrieval, and are discussed in the next section.

Until recently, structured geospatial data was authored by experts in the field, who mainly worked for national mapping agencies (NMAs), scientific organisations, or foundations. However, given the ubiquity of Web 2.0 and accurate GPS devices, web users effectively function as sensors and producers of data. Collaborative user generated content (e.g. Wikipedia) and volunteered geographic information (Goodchild, 2007) flooded the Geospatial Web. This not only led to new web-based collaborative products like GeoNames and Openstreetmap, but also spurred the growth of already existing concepts and tools such as Public Participation GIS (PPGIS) (Carver et al., 2000) dealing with the production, collection and use of spatial knowledge, through fast and free Web mapping applications. Public Participation GIS (PPGIS) (Aberley and Sieber, 2002), later, more precisely termed Participatory GIS (PGIS) (Rambaldi et al., 2006), Community integrated GIS (Elmes et al., 2005) and Participatory 3D modelling (P3DM) (Rambaldi and Callosa-Tarr, 2002) are some examples.

1.6.3 Geographic information retrieval

As discussed earlier, the web contains large amounts of unstructured geospatial data. Special methods are needed to retrieve and extract information from such large data sets in order to make sense of them. To this end, it is possible to apply and extend traditional methods from the field of Information Retrieval (IR) to unstructured geospatial data. IR can be defined as "finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" (Manning et al., 2008). Geographic Information Retrieval (GIR) is seen as a special case of traditional IR, with a focus on georeferenced content along with indexing and retrieval methods in a spatial and geographic context. GIR was first discussed by Larson (1996) in the context of Digital Libraries. Purves and Jones (2006) define GIR "as the provision of facilities to retrieve and relevance rank documents or other resources from an unstructured or partially structured collection on the basis of queries specifying both theme and geographic scope". They (Jones and Purves, 2008) subsequently list aspects of GIR as the following

- detecting geographic references in the form of unambiguous toponyms
- interpretations of vague place names and spatial language
- indexing documents according to their footprint and then ranking them according to their relevance

- developing effective user interfaces and evaluation methods

These points provide a good overview of topics of research in GIR. In particular, this thesis uses methods from the first three points, which will be discussed. The first point leads us in the direction of research performed in unambiguously geoparsing and geocoding toponyms in unstructured text.

Geoparsing and Geocoding

Geoparsing and geocoding are two important steps that need to be performed for place name recognition. It is one of the primary requirements for the work discussed in this thesis. Place name recognition is a special case of the more general problem of Named Entity recognition, a widely studied topic in Natural Language Processing. It is defined as the task of recognising location names (and other information units) and identifying references to these entities (Nadeau and Sekine, 2007).

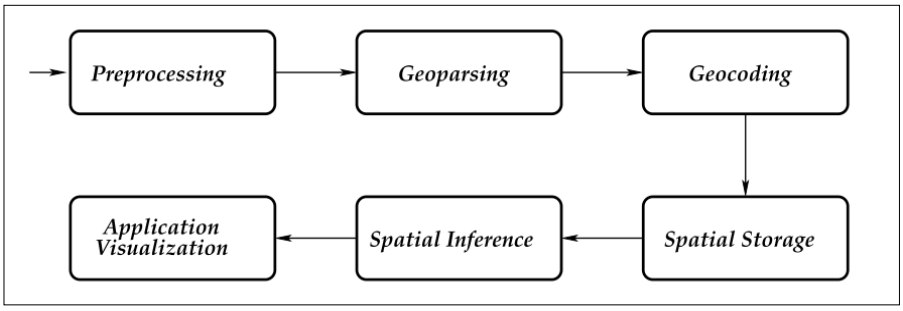


Figure 0.1 - Reference model for processing textual geographic references. (Source: Leidner and Liebermann, 2011, Page 2)

Geoparsing involves identifying place names or toponyms in a piece of text that is part of unstructured content and is commonly referred to as georecognition or toponym recognition (Leidner, 2007). As seen in Figure 0.1, given a large chunk of text, pre-processing of text is often a step that precedes geoparsing. To reduce the amount of text processing required for geoparsing, one of the first steps could be to identify commonly used words in text with the help of stop word[3] lists and/or examine the grammar and apply language rules to eliminate unwanted words (Hill, 2006; Purves, 2011). Thereafter, geoparsing can be achieved through simple gazetteer lookups

[3] List of commonly occurring words that are filtered out before processing the data. For example words like *the*, *is*, *at*, *which*, and *on* are removed and then the text is examined.

(Cunningham et al., 2002), rule based methods applied in natural language, (Mikheev et al., 1999) and machine learning (Leidner and Lieberman, 2011).

Geocoding, on the other hand, is the process of assigning unique geographic identifiers in order to gain context – usually coordinates such as latitude, longitude and altitude – to toponyms that have been extracted in the geoparsing step. This is also known as toponym resolution (Liedner, 2007). Along with the geoparsing step, this can be achieved by gazetteer lookups and also be performed using spatial expressions (Schilder et al., 2004). Silva et al. (2006) make use of an ontology of geographical concepts for extracting geographic information from large collections of web documents.

One of the main challenges with the process of geoparsing and geocoding is performing this process automatically and unambiguously, as all toponyms are not uniquely named. Toponym ambiguity is a special case of word sense ambiguity, a term commonly used in computational linguistics, for a word with more than one meaning. In the case of toponyms, this must be resolved in order to identify and ground toponyms uniquely. Amitay et al. (2004) explain that a geo/non-geo ambiguity arises if the place name has a non-geographic meaning, such as Washington, D.C. as a place vs. Washington as the name of a person, while a geo/geo ambiguity arises if there exist two distinct places with the same name, such as London, UK vs. London, Ontario. A wide variety of methods are used in dealing with toponym ambiguity, ranging from simple default rule-based methods based on, for example, population (Rauch et al., 2003; Zong et al., 2005), to methods based on exploiting toponym hierarchies (e.g. Buscaldi and Rosso, 2008), context based disambiguation (e.g. Overell and Rüger, 2008), pattern matching and cooccurrence (Li et al., 2003), or using geomorphometric characteristics (Derungs et al., 2011).

Vague place names

Vague place names and vague geographic terminology are yet another important area of research in GIR. Place names are often vague in nature. For instance, it is common for people to talk about the ‘Bernese Oberlands’ in Switzerland or ‘The Midwest’ in the USA or informal place names such as “downtown". These places exist in the minds of people and commonly occur in web documents, but it is difficult to draw a polygon around these places or express these places on a map, given that they do not exist as entries in gazetteers, mainly because their boundaries are fuzzy. Some research provides methods in identifying these kinds of regions, for e.g. Jones et al., (2008) make use of density surface modelling of other co-occurring precise toponyms to outline

vague place names. On the other hand, Montello et al., (2003) asked pedestrians to draw the location of what they thought was "downtown". Vague spatial language on the other hand refers to descriptions that humans normally communicate in, for example 'near', 'far', 'beside', 'walking distance', 'north of', 'adjacent' etc. Although it is difficult to model what these phrases mean, there is research in this direction that makes use of containment relations in addresses and fuzzy spatial reasoning to model these relations (Robinson, 2000; Schockaert et al., 2008). Similarly Hall and Jones (2008) make use of field-based models for representing vague regions and present a study for the 'north of' spatial relation.

Spatial indexing

While mining information from the web, a method of spatial or textual indexing is necessary to index and retrieve these documents accurately and effectively. This is important given that large amounts of information on the web refer to geographic locations in the form of toponyms, addresses and postal codes among others. Index structures can be put into two groups; one that makes use of individual index structures by using separate ones for textual and spatial parts and one that make use of hybrid index structures by combining the textual and spatial indexes. Zhou et al. (2005) make use of an inverted file and R* tree hybrid index for location based web search. Vaid et al., (2005) discuss three methods of indexes used along with an inverted file and perform experiments using both individual and hybrid indexes. They discuss a form of pure text indexing which was used in the Spatially Aware Search Engine for Information Retrieval on the Internet project (SPIRIT) (Purves et al., 2007). Similarly Khodaei et al. (2012) make use of a hybrid index structure to handle and rank point-based indexing of web documents in an integrated and efficient manner. While indexing methods improve performance, an efficient GIR system requires methods to rank relevant documents based on the query. Documents are typically ranked based on their textual and spatial relevance, with the former calculated based on the number of textual terms in the document and the latter calculated based on the digital footprint of the document. In the SPIRIT (Purves et al., 2007) project, relevance is computed through a footprint similarity score between the query and document footprints. Kreveld et al. (2005) discuss methods that take both spatial and textual scores into consideration, resulting in a combined score for documents.

1.7 Exploiting and extracting information from User-Generated Content

User-Generated Content (UGC) is produced on the Web and is available in huge volumes that are increasing every year. Its increase could be attributed to the ubiquitous presence of smartphones, tablets, mobiles, and other handhelds, as well as the Web 2.0 platform. The Web 2.0 platform, as discussed earlier, encourages collaboration and sharing of information through video and photo sharing websites, social networking websites, live feeds, review websites, and others. Any contribution from a user on the internet in the form of blogs, reviews, feeds, pictures, videos, or podcasts is UGC. UGC with a geographic reference, such as coordinates, is often known as Volunteered Geographic Information(VGI). Users are often able to overlay information such as running tracks or lists of their favourite restaurants on a map using simple map operations or simple map APIs provided by Google, Bing or Yahoo! Maps. The term *Neogeography* has been used to describe this phenomenon (Turner 2006) and more of it will be explained below.

Work in this thesis is focused around methods dealing with knowledge extraction from UGC and VGI in order to define regions and their attributes. The next few paragraphs discuss the phenomenon of UGC and VGI, as they comprise the main data source for the experiments and methods discussed in the following chapters.

1.7.1 Volunteered Geographic Information versus User-Generated Content

There is no generally accepted definition of UGC, but for this thesis the definition that is most applicable is: "UGC refers to media created or produced by the general public rather than by paid professionals and primarily distributed on the Internet" (Daugherty et al., 2008 page 1). Web 2.0 encourages sharing of information through easy methods and powerful infrastructure. This has resulted in large volumes of User-Generated Content (UGC) and Volunteered Geographic Information (VGI). A 2008[4] survey estimates nearly 116 million US user-generated content consumers and 82.5 million content creators, and forecasts an increase of approximately 30% by 2013. UGC is often also known as consumer-generated media (CGM) (Yap et al. 2012) or user-created content. The most popular example of UGC is Wikipedia, a free multilingual online encyclopaedia. UGC that contains location data, such as coordinates and/or toponyms, is often

[4] 2008 eMarketer survey by Paul Verna - http://www.emarketer.com/Article/Spotlight-on-UGC-Participants/1006914

known as VGI, a term coined by Goodchild (2007). It has been used frequently in GIScience during the last few years, to refer to user-generated geographic information. Seeger (2008, page 199) defines the term VGI as "geospatial data that are voluntarily created by citizens who are untrained in the disciplines of geography, cartography or related fields".

Some examples of VGI often listed are Wikimapia[5], Flickr[6], OpenStreetMap(OSM)[7], and GeoNames[8]. Wikimapia[9] is a collaborative mapping project that allow users to "mark and describe all geographic objects in the world". Along similar lines, OSM is another collaborative project aimed at creating map data that is free to use and edit (Haklay and Weber, 2008). Flickr, on the other hand, is an online service that allows sharing of photos with family, friends, and the online community (Sigurbjörnsson and van Zwol, 2008). GeoNames is a geographical names database; i.e. a gazetteer produced by web citizens. These are all collaborative projects that allow the internet user to contribute information that is shared with others. But VGI makes the underlying assumption of the user being a volunteer. In some cases, e.g. Flickr, it is not completely clear if users upload and share data with the intention of volunteering it for the use of others. For example, Ames and Naaman (2007) interviewed Flickr and ZoneTag users for their reasons behind photo capture and motivations for tagging them. It was concluded that this was done mainly to organise one's photographs and for communication with friends and family, and contextual tags were added to make the photos findable by the users and others. Platforms such as GeoNames, OpenStreetMap (OSM) and Wikimapia could be seen as examples of VGI as they involve an international and collaborative effort of gazetteer creation and mapping respectively, through volunteer effort (Goodchild 2007). Geograph[10] is a classic example of VGI, as their aim is to photograph every square kilometre of the UK and the Republic of Ireland through volunteer effort (Purves and Edwardes, 2007). Many researchers also use the term *crowd sourcing* and *neogeography* in

[5] http://wikimapia.org

[6] http://www.flickr.com/

[7] http://www.openstreetmap.org/

[8] http://www.geonames.org/

[9] http://en.wikipedia.org/wiki/WikiMapia

[10] http://www.geograph.org.uk/

connection with the above phenomenon when map users produce map data and maps and contribute to these for use by others or for private use (Turner, 2006; Ballatore et al., 2012).

But along with the benefit of large volumes of user generated geographic information, there come issues on the credibility of this information. UGC and VGI are unlike traditional authoritative geographic data that are produced mainly by national mapping agencies, governmental agencies, and professional cartographers and geographers, where credibility was generally granted as there was perceived authority due to high quality standards (Flanagin and Metzger, 2008). Spatial data quality is often examined through many aspects of quality, some of which include checking the completeness, positional accuracy, consistency, lineage, and semantic accuracy (Van Oort, 2006). Using these aspects, Haklay (2010) evaluated the quality of OSM, by comparing OSM data to Ordnance Survey (OS) (Great Britain's NMA) datasets and found that OSM data were fairly accurate for Great Britain. On similar lines Girres and Touya (2010) extend this research to the French OSM data using similar methods for quality assessment. Bishr and Janowicz (2010) propose the use of proxy measures such as informational trust and reputation to examine the quality of VGI. Ostermann and Spinsanti (2011) propose a refined workflow to automatically assess the quality of Twitter[11] based tweets, based on the original workflow of De Longueville et al. (2010), for quality assessment of VGI especially for the case of crisis management, in their case, forest fires.

Although the quality of VGI and UGC is often questioned, in some cases these data are preferred over authoritative data as they may be more up to date and time sensitive. In crisis mapping situations, for example the 2010 earthquake in Haiti, the Ushahidi[12] platform was able to collect and disseminate user-generated data very quickly through short message service (SMS), email or through the web (Roberts and Payne, 2011). Similar projects and services have been reported for disaster relief efforts such as CrisisCamp[13], Humanitarian OSM Team[14], GeoCommons[15] etc. Through information available on tweets, blogs, emails, and status updates, people and

[11] https://twitter.com/

[12] http://www.ushahidi.com/

[13] http://crisiscommons.org/

[14] http://hot.openstreetmap.org/

[15] http://geocommons.com/

organisations from all over the world are able to map and provide meaningful assistance (Zook et al., 2010).

In conclusion, UGC and VGI are of tremendous interest to the research community. In this thesis, georeferenced UGC and VGI are utilised to explore terms used by web users to describe places, points of interest and everyday activities. The work is centered on extracting and investigating information from UGC. The next section discusses specifically research that has been performed in the areas of extracting information through terms specifically from image metadata. For the remainder of this thesis, the term user-generated content or simply UGC is used as a general term to capture the phenomenon discussed above, and is intended to cover VGI as well.

1.7.2 Knowledge extraction from image metadata

It is possible to generate meaningful information from pictures posted in online communities such as Flickr, by examining their metadata. Flickr is an image and video hosting website, which in recent years, has become very popular among web users as a service for sharing and tagging their photographs. Users can annotate their pictures with tags, which is a form of metadata intended to enable users to organise and make their pictures findable. It is also possible for users to upload georeferenced images to Flickr. In Feb 2009[16], Flickr recorded 100 Million geotagged photos and estimates that the number of georeferenced photos online is between 3-4% of the total archive. To the research community, a large number of publicly available georeferenced pictures from all over the world with metadata in the form of tags describing them, along with timestamp information, are a valuable resource. Given below is a short summary of research work that has been performed using this repository.

Content analysis based on image metadata and density diagrams of images (Crandall et al., 2009) can be done by using publicly available Flickr data. Popescu and Grefenstette, (2009) and Popescu et al. (2009) collect data from large samples of Flickr images and their associated metadata (tags, geo-tags, and temporal information), and extract information about tourist trips and related tourist information. Using this they are able to predict trips for typical city tourists, by suggesting sites they are likely to visit. Similarly other researchers (Girardin et al. 2007, Girardin et al. 2008) have used these data to analyse the history of tourist presence in a city or tourist flows in cities, thereby providing an understanding of how people travel. There are large volumes of such data present,

[16] http://code.flickr.net/2009/02/04/100000000-geotagged-photos-plus/

and from a corpus of this picture data it is possible to automatically extract tourist trips (Jain et al., 2010) or analyse people's activities and movement (Jankowski et al., 2010) using georeferenced photographs. With the timestamp extracted from the photographs it is possible to carry out spatio-temporal analysis and perform a comparison of behavioural patterns of different user communities (Kisilevich et al. 2010).

Flickr is also a large repository of everyday terms that people use to describe the world around them, e.g. a place or activities performed. It is also possible to infer something about people's perception of a place, its extent and how it is described through these tags. More about people's perception of place will be discussed in the next section of the current chapter. Using these terms one can define or model vernacular places, which often have vague spatial extents. Since they are not official toponyms, gazetteers do not carry information on their spatial extents e.g. *Downtown, Alps, Mittelland* or *Midwest*. Along similar lines Hollenstein and Purves (2010) examine the use of the vernacular term *Downtown* in Flickr tags to explore how city cores across the USA are described.

1.7.3 Deriving regions from the web

Information can be harvested from various sources on the web, as described in previous sections. This information could be utilised to represent vaguely defined places, given that many vernacular regions lack a formal definition. Outside of vaguely defined places, administrative borders do not always exist in the minds of people, even in cases of places with clearly defined boundaries. It is often interesting to examine the extent of a place according to people. With UGC on the web, it is possible to learn something about people's perception of a place, its extent and how it is described. Jones et al. (2008) gathered web pages after submitting search queries to Google containing a reference to a target region, then assigned spatial coordinates to them and finally a kernel density surface was produced, representing the areas of vague places. Schockaert and Cock (2007), propose a technique using existing local search service to find fuzzy footprints of places in a neighbourhood and then attach confidence scores to these places to increase the robustness of the approach.

Pictures posted online and their metadata tell us what people are attempting to capture in these images. Thus, these sources somehow reflect *Naïve geography* (Egenhofer and Mark, 1995) or vernacular geography, and inform us about the way people think about and reason about geographic space and time. The study of differences between "absolute Euclidean spaces" and

"continuously changing notion of place" (Fisher and Unwin, 2005, page 6) through photography has triggered a lot of research in the past decade. Extracting place and event semantics, (Rattenbury et al., 2007) vague regions (e.g. Alps, Mittelland or Midwest) and their extent through Flickr photographs (Hollenstein and Purves 2010) are examples of what can be achieved with this corpus. Metadata from pictures aside, there has been a lot of research in extracting people's perception of place and vague regions from web pages (Pasley, 2008; Jones et al., 2008) and from interviews (Montello et al., 2003; Lüscher and Weibel, 2013).

Concepts of geographic place are important when examining people's perceptions. The next section therefore discusses geographic space and place and its connections to actions of people.

1.8 Geographic place and space, activities and affordances

In this part of the chapter, the thesis tries to examine previous work performed with respect to the relationship between people's interactions with place. But in order to go about this task it is first very important to first examine the concepts of space and place.

1.8.1 Geographic place and space

Cresswell (2009) explains that "*place is a meaningful site that combines location, locale and s*ense of place". Location simply refers to the "where", such as the coordinates of a certain city, e.g. 18°58′30″N 72°49′33″E, which refers to the location of Mumbai in a gazetteer. Locale refers to the way a place looks, its streets, buildings, and other visible aspects of a place, e.g. the city of Zurich has a lake, bridges over the river, a view of some churches and some cobblestoned areas. Sense of place refers to the emotions and feelings one associates with it and hence is often personal, e.g. a tourist in San Francisco has different feelings from a refugee in the same place. However, there is more than one definition of place and most researchers set out to explain place by looking at them from different perspectives. Places are represented by name, and official place names are found in gazetteers. Gazetteer entries consist of a name (possibly also including variant names), location (point, bounding box, line, polygon), and type (selected from a type scheme of categories for places/features) (Hill, 2000). In GIS, place is often represented by standard data structures; i.e. vectors and rasters. However, this method of mapping does not represent the way people think about their world (Jordan et al., 1998). Place has been described extensively by Edwardes (2007) in the context of LBS.

Space is represented as geometry, which allows one to encode the world in a purely logical way (Edwardes and Purves, 2007). It is represented absolutely using a spatial reference system or a coordinate system that could be local, regional or global. It allows positions to be indexed and organised and is the most straightforward method to link geographic information and the user (Edwardes, 2009). In Figure 0.2, Edwardes (2007) illustrates that place lies at the opposite end of a continuum of geographic perspectives from space and that VGI provides a real opportunity to describe place.

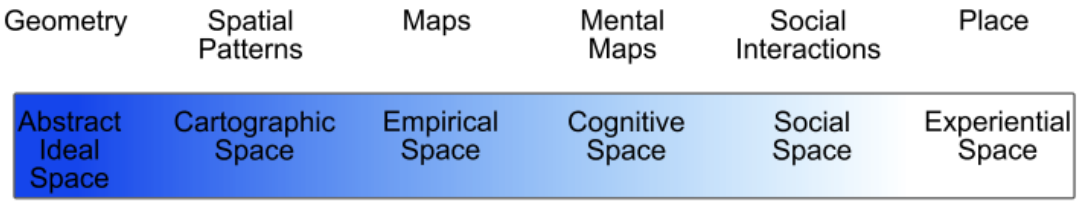


Figure 0.2 - The space-place continuum (Edwardes, 2007)

The concept of affordances is very important for this thesis, as one aspect of it is centered on what actions places afford humans. The general term was introduced by Gibson (1979), who examined how people perceive their environment around them. Affordances are what objects or things offer people to do with them, thereby creating activities for users (Jordan et al. 1998). It is difficult to define the concept of affordances, but it can be described with examples: a chair affords sitting, a park in the city affords jogging, playing, and chatting (Ostermann and Timpf, 2007). Affordances are therefore a result of the mental interpretation of things, based on people's previous knowledge and experience. Looking back at the space-place continuum by Edwardes (2007), affordances can be placed in an experiential view of space and place, because they offer a user-centered perspective (Kuhn 1996). Affordances and what places afford humans to do are discussed in detail in the next section along with work performed and methods developed in the direction of extracting people's notion of places and regions.

1.8.2 Affordances and activities

One could think of human activities as having an overall space time metric as they occur within a context of locational space coordinates, are often separated within a space metric, and occur within a time period (Golledge and Stimson, 1997; Miller, 2005). In GIScience activities are often reviewed through the concept of space-time prisms (Miller, 2004 and Yu and Shaw 2008). In the past activities have often been seen as an attribute of a place or even something that just describes

it. Tversky and Hemenway (1983), in their attempt to build a taxonomy for scene categories (e.g. indoors, outdoors), carry out experiments on how people categorise scenes using basic (e.g. school, home, beach, mountain) or preferred levels of categorisation based on perceived attributes such as activities. Jordan et al. (1998) in their discussion on an affordance-based model of place, list several aspects of place, one of them being the "actions people perform in a place". Relph (1976) states that one of the unique qualities of a place is "its power to order and to focus human intentions, experiences and actions spatially" and also a "unique instance of observable activities and functions". In the field of environmental psychology, Canter (1997) states his theory that place has four faces namely functional differentiation (related to activities), place objectives (related to individual, social and cultural aspects of place), scale of interaction, and aspects of design (physical characteristics). Edwardes (2007, page 3) believes that "places are created as centres for actions, meanings, interactions and everyday rituals" and that "place-based thinking relates to direct experiences, actions and activities conducted in space". He also explains that people see their environment as affording opportunities to engage in particular activities. Hence space and place are important dimensions of human activities. The linking of space with activities is not a new concept, as it has been discussed earlier in studies in geography and human geography (Pitzl, 2004:5).

However, some researchers believe that as transportation and communication costs are collapsing, the world is shrivelling and fragmenting, and people and activities are becoming disconnected with location (Couclelis and Getis, 2000). However, this was researched in the context of activities in urban areas, such as activities performed at home and office, where there seemed to be a growing trend in working from home, online shopping in the office, etc. These activities no longer are linked to or affected by place. However, this thesis investigates activities not performed at office, home, and urban spaces, but mainly activities that link to the characteristics of place such as landscape and terrain.

There is no standard method of gathering these activity terms in spite of there being many ways in which activities or actions performed by humans are determined. Daily activities performed can be recorded by using wearable devices that automatically detect what a person is doing (Lester et al., 2005), or through RFID tags (Buettner et al., 2009) or by simply asking people through a questionnaire (Tinsley and Eldrege, 1995). Extraction of activities particularly from UGC has also been performed. Dearman and Truong (2010) automatically extract activities from the review website Yelp. They harvest this information using Yelp's web service and make use of NLP methods to detect nouns and verbs. Purves et al. (2011) looked at verbs and nouns in Flickr tags

and analysed the most frequent terms in the collection. They then employed a voting mechanism to decide which term in the list was eligible to be an activity. Once extracted, Joshi and Luo (2008) make use of Flickr tags and metadata to compute statistical saliency of geo-tags in describing an activity or event. They do this by categorizing tags into event or activity classes using visual detectors that perform pure visual event and activity recognition.

Overall Research Methodology

This chapter discusses the research gaps that were observed after having studied the already existing infrastructure and background in the preceding chapter. It summarises gaps that were identified and explains the reasons behind choosing the research questions discussed in Chapter 1.

1.9 The coverage of the Geospatial web

The use of web content, both in the form of unstructured text, as well as objects with explicit georeferencing, is an increasingly popular way of exploring a wide range of geographic questions (Egenhofer, 2002; Leidner and Lieberman, 2011). However, it is very unlikely that web content is evenly distributed in space, and studies which seek to draw conclusions based on, for example variations in density, must first estimate the underlying density of the collection of interest. Implicit assumptions about homogeneity of coverage can be misleading. Pasley et al. (2008) set out to explore how web coverage varied for different forms of social media in the UK, correlating coverage of a variety of sites with overall web coverage and population.

What is missing is the focus on examining the coverage in unstructured textual information such as web documents, rather than in structured databases. Although previous work has shown that web coverage is, unsurprisingly, not homogeneous (Pasley et al., 2008 and Venkateswaran, 2010), only a few researchers address the issue of how it varies, beyond obvious relationships to population. For example, Crandall et al. (2009) show a density map (Figure 0.1) of geotagged Flickr images from all over the world. From this map, one might hypothesise that density of Flickr images correlates with population, but this is only true for the Europe and North America. There exist other factors such as internet connectivity, popularity of Flickr as a social media service, popularity of a given place due to tourism, or some other explanatory variable. With such hypotheses as a starting point, part of this thesis examines the web coverage and correlation between web coverage and possible predictor variables that could be used to explain it. As an application domain and study area, web coverage is examined through tourism in Switzerland. Tourism is an important contributor to the Swiss economy, and is often used as a prototypical application for the utilisation of

web content. Switzerland is a multilingual country and is frequented by tourists speaking many different languages. Therefore, to complete the picture not only geographic web coverage, but also variations therein as a function of language will be examined.

A few researchers have already realised the problems of uneven data coverage. Graham et al. (2012) discuss the digital divide that is caused due to Internet access; i.e. the digital divide in the geography of the internet by considering the raw number of internet users in each country as well as the percentage of the population with internet access. They later examine the georeferenced tweets produced by Twitter users all over the world and plot a spatial tree map (Figure 0.2). This map clearly shows the inequality in the geography of content. Li et

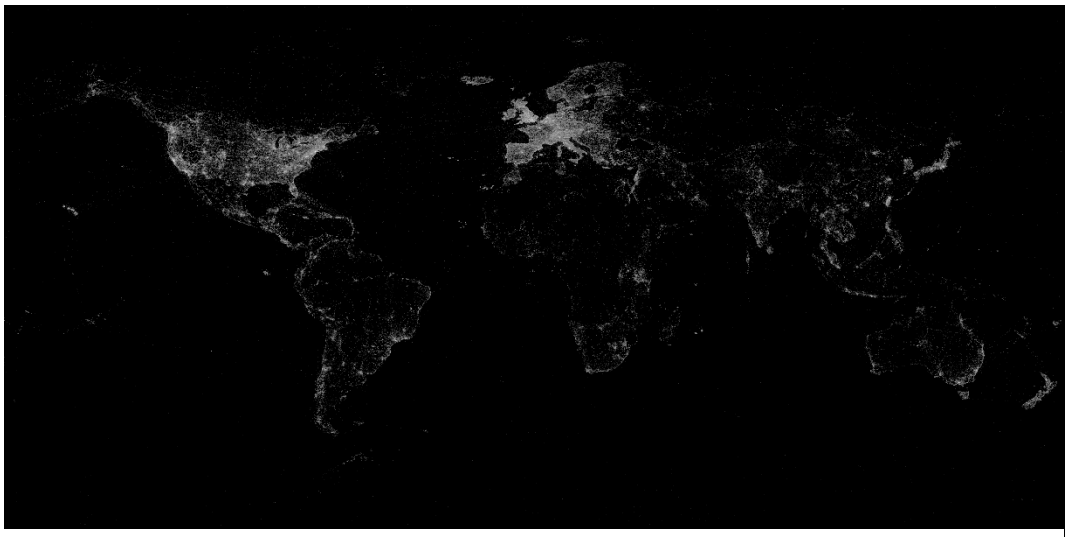


Figure 0.1 - World heat map of geotagged photos from Flickr (Crandall et al., 2009)

al. (2013) use georeferenced Twitter and Flickr data to derive patterns rather than using only one of them, as they acknowledge that there is uneven distribution of the data generated in social media and the nature of such data has to be understood and used appropriately. All the above work suggests that the content that exists on the web is not homogeneous and varies due to a number of reasons.

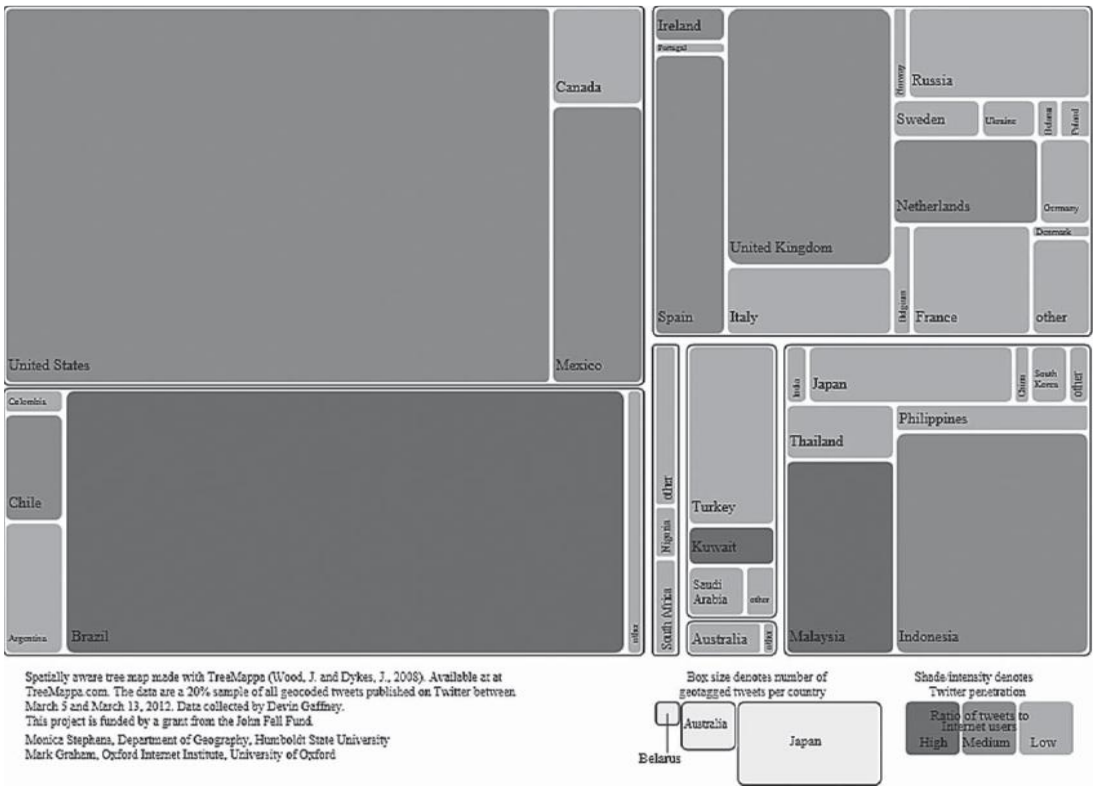


Figure 0.2 - Spatially aware treemap representing number of Twitter tweets (Graham et al., 2013)

Research approach

Pasley et al., (2008) used counts as a proxy for coverage, and therefore density, by retrieving the total number of occurrences of documents with a given toponym from a search engine index via an API (Application Programming Interface). The same approach is used to examine coverage, in order to answer the research questions. This is done by counting the individual number of web pages returned by a search engine based on tourism search queries containing toponyms (place names) obtained from gazetteers. The toponyms for the search phrase were taken from the following three datasets: SwissNames, the Tele Atlas Points of Interest (POI) dataset, and the GeoNames gazetteer dataset, in order to reduce coverage bias from the gazetteer itself.

1.10 Determining and describing locations and related activities from User Generated Content (UGC)

In the past, place and geography of place have been described with the help of UGC, both text and images. In previous research, georeferenced data on the web has been automatically extracted and annotated (Purves et al.; 2008, 2010; Rattenbury et al., 2007) and this information was used to describe places, for example through geographic elements, qualities and activities (Purves and Edwardes, 2007; Edwardes and Purves, 2007; Purves et al., 2011). Hence, building from previous research, part of this thesis aims to describe what actions people think they are performing by examining online photos. People's actions are an important aspect of place (Jordan et al., 1998) and in this thesis actions are measured through activity terms extracted from UGC. Although there has been considerable research in extracting activity terms and building ontologies of activities, there has not been much research in the area of linking activity terms to place. Examining what people are photographing and what constitutes, in their minds, a tourist related activity for a given place, is one of the sub objectives of this thesis. However, this goal requires a repository of georeferenced pictures that contain metadata such as descriptions of the pictures. Extraction of this information from user generated content is possible from Flickr images, especially through social media that contain explicitly georeferenced information.

Shatford (1986) identified and classified the kinds of subjects a picture may have, this is summarised in Table 0.1. This categorisation discusses four facets of image classification; *Who, What, When,* and *Where.* Each of these basic facets were subdivided into aspects based on 'Of in the specific sense', 'Of in the generic sense', and 'About'. This categorisation has been extensively used to explore how images are described in UGC and also used for categorising images based on their descriptions. Purves et al. (2008) emphasize the importance of the 'Specific Of' and 'Generic Of' elements of the 'Where' facet in generating suitable annotations for images. Toponyms are typical examples of the 'Specific Of', while geographic kinds, such as a church, hill or lake illustrate the 'Generic Of'.

Research approach 1

One of the sub-objectives of this thesis is to relate a location to actions through pictures; i.e. to link the *What* and *Where* aspects in the matrix shown Table 0.1. Although there has been

extensive work in automatically extracting locations through UGC and especially images, there is a need for a simple repeatable method to automatically outline locations based on people's perceptions of place, through descriptions provided in the Flickr images. One can argue that every georeferenced picture provides a location in the form of coordinates, and information on administrative boundaries could be procured from any gazetteer, but using this information defeats the objective of examining people's perception, as boundaries and locations are also factors of perception. The next proposal is to link location to people's actions and what activities people perform or perceive is being performed in a picture. Referring back to the matrix, the proposal is to link the *What* and the *Where* aspects through images.

Table 0.1 - Shatford's (2008) Faceted Classification of the Subjects of Pictures

Facets	Specific Of	Generic Of	About
Who?	Individually named persons, animals, things ...	Kinds of persons, animals, things	Mythical beings, abstraction manifested or symbolised by objects or beings
What?	Individually named events	Actions, conditions	Emotions, Abstractions manifested by actions
Where?	Individually named geographic location	Kind of place geographic or architectural	Places symbolised, abstractions manifest by locale
When?	Linear time; dates or periods	Cyclical time; seasons, time of day	Emotions or abstraction symbolised by or manifested by time

An example of people's perception of place is discussed in Figure 0.3 (Hollenstein, 2008) shows extracted points of georeferenced Flickr images taken on one day. These points contain terms such as 'city' and 'Zurich', used to describe the central area of Zurich. Thereafter, a kernel density estimate has been calculated. This image shows an example of a method used to outline vague footprints and is the starting point of outlining areas.

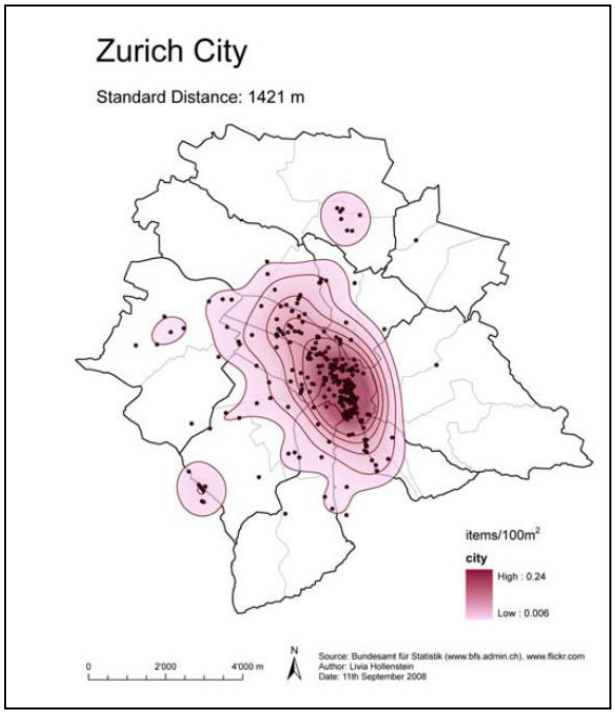


Figure 0.3 - Kernel density estimate with isolines showing the footprint of Flickr photos with the tags Zurich and city (taken from Hollenstein, 2008).

Research approach 2

Activity and location are linked through the concept of affordance; we examine the kind of activities afforded by a certain location. Affordances are what objects or things offer people to do with them, thereby creating activities for users (Jordan et al. 1998). For example, a chair affords sitting and a park in the city affords jogging, playing, and chatting. (Ostermann and Timpf, 2007). Similarly, a location in the mountains may afford activities of hiking and trekking, and the infrastructure of a particular location may afford certain other activities. The focus in Chapter 6 will be on examining this.

The area of activity theory relates to the theory, measurement, and analysis of how people organise their activities in space and time. It is a framework whose foundational concept is activity, which is a "purposeful, transformative, and developing interaction between actors

("subjects") and the world ("objects")" (Kaptelinin, 2012). How factors like transportation, communication and settlement systems are related to activity patterns and in turn their influence on the organisation of activities (Miller, 2005). Miller (2005) summarises the basic components of activity theory done by Wang and Cheng (2001) shown in Table 0.2. This table is used in 0 to study the interaction between activities and their behaviour.

Table 0.2 - Table showing the basic components of activity theory (Wang and Cheng, 2001; Miller, 2005)

Entity	Definition
Activity	The main purpose carried out at a location, including waiting time before and afterwards. Activities can be classified into different types depending on purpose
Activity frequency	The number of time the activity occurs during a given time period
Activity destination	The location where the activity occurs
Trip	Movement between two activity destinations
Transport mode(s)	Methods of conveyance used to perform a trip
Activity program	Set of activities to be performed within a given time period
Activity schedule	The planned ordering of activities in space and time within a given time period
Activity pattern	The activities in space and time within a given time period
Activity space	A composite of the locations where an individual conducts routine activities
Physical environment	Spatial configuration of the activity destinations and transportation services between these destinations
Institutional environment	Set of formal rules the regulate the individual's activities in space and time (e.g. store hours, working hours)

Of Web counts: Geographic coverage and linguistic differences[17]

1.11 Introduction

The use of web content, both in the form of unstructured text as well as objects with explicit georeferencing, is an increasingly popular way of exploring a wide range of geographic questions (Egenhofer, 2002; Leidner and Lieberman, 2011; Jones and Purves, 2008 and Purves 2011). However, as every geographer knows it is very unlikely that web content is evenly distributed in space, and studies which seek to draw conclusions based on, for example, variations in density must first estimate the underlying density of the collection of interest. Implicit assumptions about homogeneity of coverage can be misleading, as shown by Pasley et al. (2008) in a study exploring variations in web coverage across different forms of social media in the UK, where they examined the correlation of the coverage as seen in a variety of sites with overall web coverage and population.

The main contribution of this chapter is the use of the web counting method (Kilgarriff and Grefenstette, 2003) in order to create a repeatable method for examining the web coverage and its variations caused due to language. This is done by counting the individual number of web pages that are returned by a search engine based on tourism search queries containing toponyms (i.e., place names), obtained from gazetteers. This chapter discusses web coverage in detail exploring how it is affected by different influencing factors. Importantly, this chapter focuses on examining the coverage in unstructured textual information such as web documents, rather than in structured databases. Visualisations are presented through graphs and coverage maps and geo-processing methods are used to analyse the results through correlations to basic potential influencing factors of web coverage. For the remainder of this thesis, the term web counts refer to the number of web documents that exist in a web text corpus for a given toponym. However, this web text corpus is not the complete collection of web documents on the internet; instead it refers to the collection of documents that are matched to an input query by a web search engine. As an application domain

[17] This chapter has been published and has been extracted from Venkateswaran et al., (2014)

and study area, tourism in Switzerland is used as discussed in the earlier chapters, since tourism is an important factor in the Swiss economy, and since tourism is often used as a prototypical application for the utilisation of web content. Switzerland is a multilingual country and is frequented by tourists speaking many different languages. Therefore, to complete the picture not only the geographic web coverage, but its variation as a function of language is examined. Ad hoc tourist information is readily available on the web in the form of pages that contain news, lists, catalogues, reviews, blogs and multimedia content related to activities targeted at particular regions. Although previous work has shown that web coverage is indeed not homogeneous (Pasley et al., 2008 and Venkateswaran, 2010), little work has addressed the issue of how it varies, beyond obvious relationships to population. For example, Crandall et al. (2009) show a density map of geotagged Flickr images from all over the world. From this map, one might hypothesise that density of Flickr images correlates with population, internet connectivity, popularity of Flickr as a social media service, popularity of a given place due to tourism, or some other explanatory variable. With such hypotheses as a starting point, the correlation between web coverage and possible predictor variables that could be used to explain it, are examined. Detailed local knowledge is necessary to analyse and discuss the spatial patterns and geographic relationships identified in work of this nature. Once again, since our case study focuses on tourism, possible variables bear a relation to tourism and related influences such as language. The key questions driving this research are therefore:

1. How does the geographic distribution of web coverage for tourism-related themes vary across Switzerland?

2. Are there any differences in web coverage distribution for different languages and gazetteer datasets?

3. How do typical factors such as population and touristic popularity of a place affect web coverage?

Our underlying motivation is to develop a simple, repeatable method which allows us to explore web coverage. Such maps of coverage can then be used as baselines to explore variation in coverage (either in time or space), rather than simply assuming either homogeneity of coverage or that coverage simply varies as a function of population.

1.12 Data and Methods

1.12.1 Approach

This section introduces the methods used to establish the geographic and linguistic web coverage for tourism in Switzerland. From previous work performed (Venkateswaran, 2010) there are hints that web content is linked to and varies with language. Since the coverage problem focuses on tourism-related themes in Switzerland, it is essential to first understand the linguistic background of Switzerland. Switzerland has four official languages: German, French, Italian and Romansh. The number of Romansh speakers, according to the Swiss Federal Office of Statistics, was approximately 35,095 (2000 Swiss census), which makes up 0.5% of the population of Switzerland. Similarly, the number of native speakers is approximately 64% for German (all dialects), 20% for French and 6.5 % for Italian. Given the proportionally low number of Romansh speakers, English was selected over Romansh, as this study focused on tourism and English is an important language in that respect. Web counts were therefore examined in German, French, Italian, and English.

In order to gather web counts, the following trigger phrases: <"Toponym" Schweiz tourismus>, <"Toponym" Suisse tourisme>, <"Toponym" Svizzera turismo> and <"Toponym" Switzerland tourism> were used. These phrases were made up of a toponym, followed by keywords for Switzerland and tourism in the four different languages. The toponyms were selected from several datasets that contained names of populated places. An example of a search query is thus <"La Chaux-de-Fonds" Switzerland tourism>. Toponyms were placed in quotes so that only exact matches were found, in particular for toponyms which are made up of multiple words. The phrases selected resulted from initial testing with a combination of the toponyms with canton[18] names, country and tourism related terms such as ‘attractions’, ‘places to visit’ etc. The country along with the keyword ‘tourism’ seemed to yield the highest web counts. Also, work done by Hollenstein and Purves (2010), reports that tourists are more likely to tag photographs on Flickr as a combination of the place name and country rather than with the state, county, or canton.

In the case of the point of interest (POI) data, specifically related to tourism, the word ‘tourism'and its translations in the three other languages were omitted from the search phrase. This is because

[18] Switzerland is a federal republic made up of 26 cantons.

most POIs were typical tourist locations, hence it could be assumed that the toponym was directly related to tourism.

To determine the number of hits (denoted as web counts in the remainder of the chapter) the Yahoo! Search BOSS API was used for the above sets of toponyms along with tourism phrases in the four languages. The API has a wide variety of parameters that can be supplied and that may influence results. For instance, the type of the web content can be specified using the type parameter and this includes specifying the format of the documents that match the search query, for instance html, text, pdf, doc, etc. Since the main aim was to study the aggregate coverage, it was concluded that the format of the web content did not matter and that any web page that contained these terms was candidate towards contributing to the web count. Another instance is the query operator. Boolean operators like 'AND' and 'OR' can be used to combine query words, and hence could be used in the search phrases that were formed (discussed above). However, there was no significant change in the web count with or without the AND operator, hence it was not made use of. Furthermore, the search was not restricted to the top level domain '.ch', since many tourism websites are hosted under '.com'. Finally, as locale the default 'en-us' was used, since preliminary experiments had shown that many tourism websites use this locale rather than the locale of the local language spoken (e.g., 'de-ch', 'fr-ch' etc.).

These counts were extracted in February 2010 and this cache of counts has been analysed further in the research. The API returns two values called *totalhits* and *deephits* respectively. Both these values are approximate counts of the number of web documents that exist as, firstly, the Yahoo! Search BOSS API returns only a smaller proportion or a snapshot of the web, instead of all web documents and, secondly, the number of hits returned is an approximation based on proprietary code. *Totalhits* does not contain duplicates while *deephits* reflects duplicate documents and all documents from a host. Hence, *totalhits* was selected as the web count for the study.

1.12.2 Toponym data

The toponyms for the search phrase were taken from the following datasets: SwissNames, Tele Atlas Points of Interest (POI) dataset and the GeoNames gazetteer dataset. SwissNames is provided by the Swiss Federal Office of Topography (swisstopo). The dataset contains 155,571 place names in 62 categories shown on the swisstopo 1:25,000 map, and contains other essential pieces of information such as coordinates, altitude, 'Gemeinde' (town) name and canton name. The POI dataset was provided by Tele Atlas BV 2010. It contains 54,912 points of interest in 50

categories. The POIs are attributed with important information including coordinates, name, address and other details. The GeoNames gazetteer is provided online by www.geonames.org. The data for Switzerland contain 20,726 place names in 107 categories, also known as feature classes. One of the highlights of the dataset is that along with *placenames*, it also lists *asciinames* and *alternatenames*. The *asciinames* restrict spellings to only ASCII letters, while *alternatenames* spell out the place name in a number of other languages. Modifications of spelling in Section 1.12.2 discusses how this additional information was utilised.

The above selection of datasets covers three different types of data sources: Topographic data by a national mapping agency, POI data from a commercial data provider, and, to some extent in the case of GeoNames, user generated geographic content. In the following sections, the analyses that make use of the above datasets are described.

Settlements from SwissNames

From SwissNames, all the toponyms of populated places were selected: cities, towns, villages and settlements as shown in Table 0.1. This toponym set contained 7,949 populated places in Switzerland. Out of the 7,949 records, 1,704 places were eliminated because of geo/non-geo ambiguities that caused the counts to be artificially high (cf. Section 1.12.3 for more detail on ambiguities). Following this, web counting was achieved using the approach described above.

Tourist destinations from Tele Atlas POI

From this dataset, a list of 787 tourist destinations were extracted from the Tele Atlas database, by filtering towns or points of interest that were explicitly marked ‘Important Tourist Attraction’. The web counting approach described above was then performed. No ambiguities were identified, in contrast to the previous analysis with SwissNames.

Populated places from GeoNames

For the third analysis, names of populated places were extracted from the GeoNames.org gazetteer. As discussed above the gazetteer provided information on toponyms, their corresponding feature codes and population. Among all the toponyms, only toponyms with feature code ‘PPL’ and ‘PPLA’ were chosen. In GeoNames, PPL is a populated place and is defined as "a city, town, village, or other agglomeration of buildings where people live and work", while PPLA is a seat of a first-order administrative division. Other populated toponym categories like PPLA2, PPLC, PPLL exist but were not considered for the study, as they were too small (population-wise) or

already included in PPL and PPLA. 4,337 entries were initially selected, of which 412 were deleted due to geo/geo and geo/non-geo ambiguities and another 277 were aggregated (by the method discussed in Section 5) and then deleted due to repetitions.

Table 0.1 SwissNames list of populated places that were selected for the experiment

SwissNames code	Explanation
HGemeinde	city > 50,000 inhabitants
GGemeinde	city 10,000 - 50,000 inhabitants
MGemeinde	town 2000 - 10,000 inhabitants
KGemeinde	village < 2000 inhabitants
GOrtschaft	large settlement > 2000 inhabitants
MOrtschaft	middle settlement < 2000 inhabitants
KOrtschaft	small settlement 50 - 100 inhabitants

Table 0.2 - SwissNames list of Mountain features, in German and English

Berge	Mountains
Massiv	important mountain range
HGipfel	main Alpine peak
GGipfel	minor Alpine peak
KGipfel	small peak
Grat	Ridge
Fels	Rock

Mountains from SwissNames

A similar experiment was carried out for mountains. SwissNames contains 7 categories of mountains (Table 0.2). From this table 'Grat' and 'Fels' were removed, assuming that they do not play an important role in places for tourism and sports as compared to the ones selected. From the rest the first 552 peaks were selected and the same experiment was repeated as above. Similarly, for Tele Atlas POI all records of 'Mountain Peak' were selected and the same experiment was carried out. For this experiment all ambiguities were deleted and so were all the records with zero

counts. The resulting web counts of mountains in Switzerland showed a similar trend as with the previous experiments.

Modifications to toponyms in SwissNames and GeoNames

After the four sets of toponyms were selected, some translations and changes in the spelling of toponyms were made. These modifications were performed on the settlements (SwissNames) and populated places (GeoNames), as the web counts could be slightly skewed or biased for several reasons, such as:

- The datasets contain toponyms that are in the local language. For instance, the towns in the French speaking part of Switzerland are in French (e.g. Genève) and towns in the German speaking part of Switzerland are in German (e.g. Zürich). This skews the search results and in turn the web counts, as the local name may not be used in a website of a different language, hence not reflecting the real nature of the coverage. Therefore, these web counts were also examined after translating the toponyms to the particular language of examination (e.g. Geneva for English and Zurigo for Italian). All the translated names of the toponyms were extracted from Wikipedia using WikAPIdia19.
- Occurrences of diacritics such as ‘ö’, ‘é’, ‘è’, etc. in a toponym are highly language specific. The content on the web in a particular language often does not contain toponyms with the special characters of another language. For instance, ‘Zürich’ is spelt as ‘Zurich’ in English and French, causing the counts to be skewed, as the search phrase <"Zürich" Switzerland tourism> does not appear as frequently as <"Zurich" Switzerland tourism>, and similarly in French. A preliminary examination caused the number of counts to drastically increase to 111,139 for ‘Zurich’, as compared to 28,227 for ‘Zürich’ in English. Hence, on the basis of this observation, web counts were also examined by considering the toponyms in the ASCII form, after replacing any non-ASCII character with its respective ASCII character (for example ‘ü’ with ‘u’, ‘è’ with ‘e’ etc.).
- Some toponyms in Switzerland are spelt with another ‘e’ to replace the umlaut diacritic (¨) in the German spelling (Table 0.3). Thus ‘ü’ is replaced by ‘ue’, and ‘ä’ is replaced by ‘ae’. This was also applied to the toponym set.

[19] http://collablab.northwestern.edu/wikapidia_api/Wikapidia/Home.html

If a toponym did not have a translation or diacritic, then the (unchanged) web count for the original toponym name was considered.

Table 0.4 shows a case by case example of what is discussed above and how web counts change depending on whether the toponym is in the local language, taken in ASCII format, spelling changed, or translated.

Table 0.3 - Examples of spelling changed toponyms (only for English)

Name	Changed spelling
Zürich	Zuerich
Graubünden	Graubuenden
Grächen	Graechen

Table 0.4 - Comparison between original, ASCII-converted, spelling changed and translated toponyms

Original counts	Higher counts	What was higher?
Glarus Suisse tourisme ~3000	Glaris Suisse tourisme ~20000	Translated toponym into French (from German)
Neuchâtel Switzerland tourism ~20000	Neuchatel Switzerland tourism ~30000	ASCII toponyms (instead of French spelling)
Zürich Switzerland tourism ~30000	Zuerich Switzerland tourism ~35000	Toponym without diacritic (from German)

Having carried out all of these operations, a final set of nine toponyms datasets was generated (Table 0.5). For each of these nine toponym sets, the web counts were generated in the four languages, amounting to a total of 36 individual runs. More ambiguities, both geo/geo and geo/non geo, were found. However, they were proportionally much less numerous than the ones found earlier and thus were removed from our experiment.

Table 0.5 - Final list of toponym sets

Analyses	Sets containing
Settlements from SwissNames	1) Original toponyms 2) translated toponyms 3) ASCII toponyms 4) toponyms without diacritic
Tourist destinations from Tele Atlas POI	5) Original toponyms
Populated places from GeoNames	6) Original toponyms 7) translated toponyms 8) ASCII toponyms 9) toponyms without diacritic

1.12.3 Toponym ambiguities

Geo/geo ambiguous toponyms were disambiguated using a simple, but effective metric; population, which results in a one sense per discourse representation (Rauch et al., 2003). Table 0.6 shows an example, where 'Aesch' is treated as a geo/geo ambiguity and the duplicate entries were deleted by the procedure explained above. Table 0.6 also shows another effect, visible in the last column. Occasionally, there were toponyms that shared the same name but had different web counts. While this may seem surprising, this difference can be attributed to cache updates that might happen on Yahoo! at any point of a processing run. Hence, as the final count, the highest web count was selected for the set of toponyms that had a common name. The above strategy should work in most cases, but if the toponym is an obvious tourist destination with few permanent inhabitants, this approach may not work.

Table 0.6 - Ambiguities in toponyms. Numbers in bold typeface denote final values chosen as explained above.

Toponym	Coordinates	Population	German web counts
Aesch	47.47104, 7.5973	**10138**	3791
Aesch	47.26667, 8.25	911	3791
Aesch	46.88333, 8.8	0	**3988**

In the case of geo/non-geo ambiguities, a stop word lists of common words and commonly used geographic terms such as 'berg' (mountain in German), 'stein' (stone in German) etc., was used and was made in four languages. Toponyms with these names were automatically deleted and not examined. Also, simple methods such as comparing the web counts to population were used and several toponyms with an extremely high web count but a very low population count were identified. With the help of local knowledge, it was concluded that many of these toponyms were

geo/non-geo ambiguities, and like the previous ones they were deleted from our list. Finally, after a manual inspection of the list of the top 100 web all the geo/non-geo ambiguities identified for all four languages and thereby deleted. Table 0.7 shows some examples of typically occurring toponym ambiguities, along with the number of times they appear in the SwissNames dataset. Table 0.8 shows the pre-filtered top 10 Web counts in four languages. It is clear that high web counts are dominated by ambiguous uses of toponyms, which typically do not refer to locations, and thus filtering the web counts as described above appears to be a sensible strategy.

Table 0.7 - Ambiguities in toponyms. Language wise typically occurring top 5 toponym ambiguities.

Toponym (German)	No. of times	Meaning in English	Toponym (French)	No. of times	Meaning in English	Toponym (Italian)	No. of times	Meaning in English
Alle	1	All	Au	9	To	Del	1	The
Platz	2	Place	Nord	2	North	Alle	1	To
Markt	1	Market	Plan	2	Map	Stampa	1	Print
Bild	2	Picture	Mon	1	Mine	Nord	2	North
Berg	11	Mountain	Premier	1	First	Valle	1	Valley

Table 0.8 - Top 10 language wise Web counts pre-filtered for toponym ambiguities

Toponym (German)	Web count	Toponym (English)	Web count	Toponym (French)	Web count	Toponym (Italian)	Web count
Alle	404649	First	1252161	Au	1219563	Del	392964
Platz	241955	Costa	1126839	Nord	682306	Alle	214242
Markt	229388	Full	821796	Plan	635139	Stampa	149994
Bild	211901	Sales	582168	Mon	454508	Nord	120429
Berg	210786	Plan	548689	Premier	381920	Valle	97877
Buch	183385	Far	413484	Provence	329040	Costa	93269
Ins	154673	Bissau	375120	Rue	286517	Strada	79388
Schutz	151891	Seen	314905	Font	232269	Far	71711
Plan	140621	Play	308726	Champagne	209970	Piazza	70948
Bad	131255	Says	291378	Tavers	195906	Isola	67997

1.13 Geographic web coverage

This section discusses the results and analyses related to Research Question 1, seeking to establish the geographic web coverage for tourism-related themes in Switzerland. Figure 0.1 shows the resulting web counts for Swiss toponyms. The graph is a plot of the web counts vs. places, and is plotted on a logarithmic scale with colours reflecting different gazetteer data sources: SwissNames (blue lines), Tele Atlas POI (black), and GeoNames (red). In each colour, the four different lines indicate the four different languages that were chosen for this study. Web counts were sorted individually for each language in decreasing order. That is, the sorting order differs between languages, and thus the graphs suggest the trends and the frequency distribution of the web counts over all toponyms, rather than the specific web counts per individual toponym.

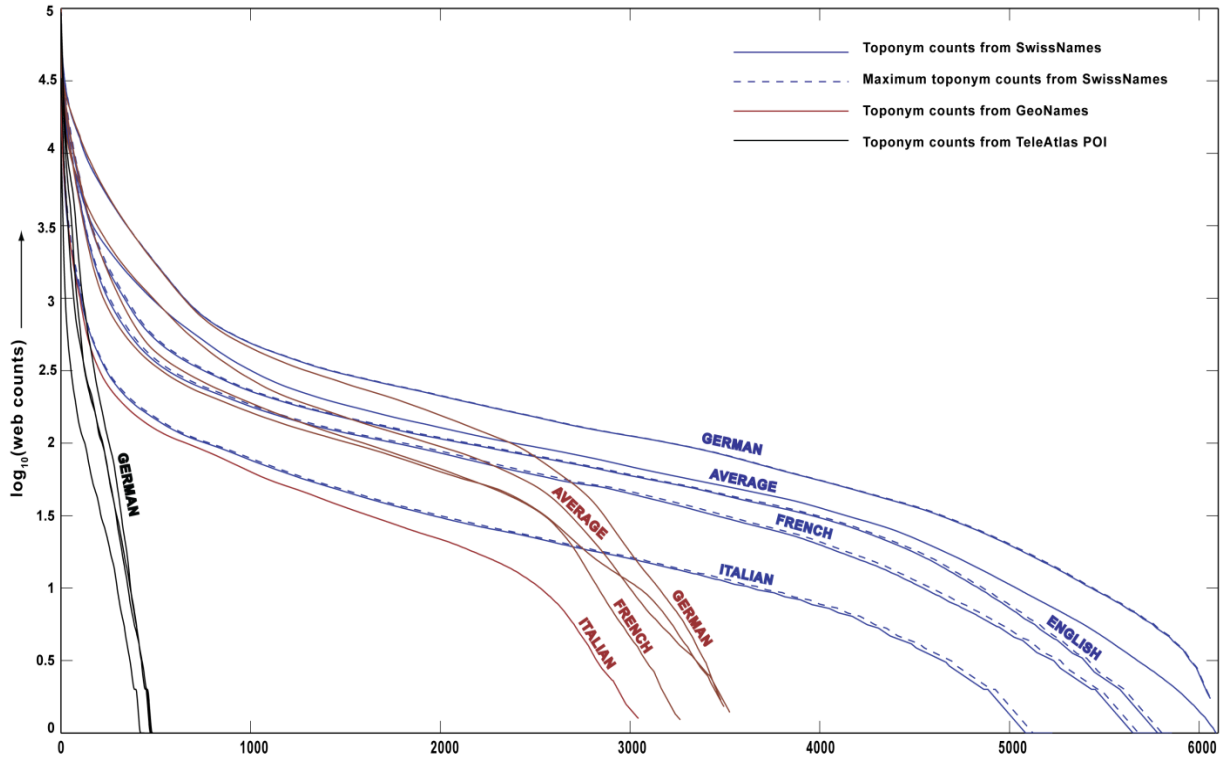


Figure 0.1 - Plot of web counts vs. places (tourist attractions), plotted on a logarithmic scale with colours reflecting different gazetteer data sources: SwissNames (blue), Tele Atlas POI (black), and GeoNames (red). Dashed line denotes maximum toponym counts res resulting in spelling modifications made as explained in the earlier section. In each colour, the four different lines indicate the four different languages that were chosen for this study

Looking at the graph in Figure 0.1 the general trend seems to be that German constantly has the highest counts and it seems to have the highest number of counts for many places, as compared to English, French or Italian. This reflects the dominance of German as the most widely spoken language in Switzerland. The counts for Italian, on the other hand, are lowest, again in line with the observation that Italian is less frequently spoken in Switzerland than German and French.

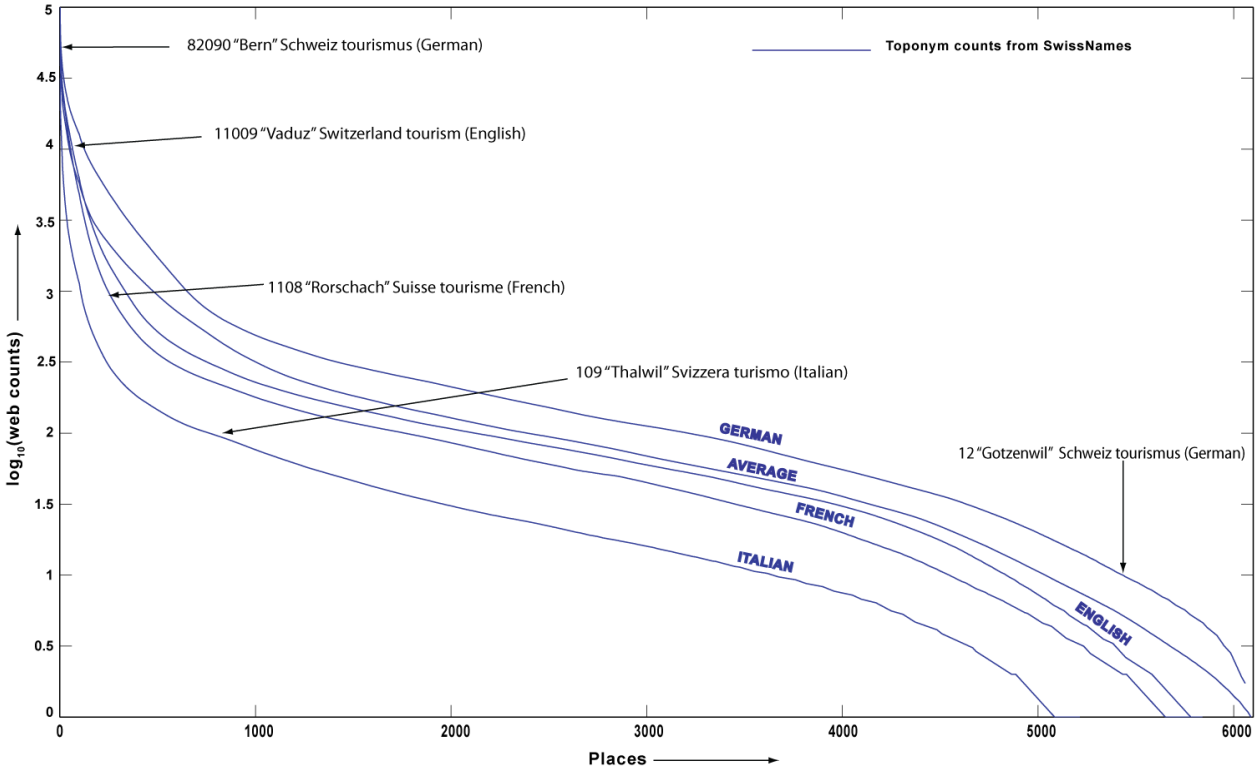


Figure 0.2 - Selection of SwissNames from Figure 4.1 with place names and its approximate value of the web count for selected places.

Counts for Tele Atlas POI data were lowest. This happened despite the fact that the word 'tourism' (and its translations) was omitted from the search phrase used for this gazetteer data set, resulting in a less restricted search. This result suggests that the cumulative tourism web content in Switzerland is greater for individual cities than for specific tourism attractions. In other words, city names are typically used when referring to tourism rather than more specific names. The graph seen in Figure 0.2 represents a selection from the graph seen in Figure 0.1, focusing on the results for the SwissNames gazetteer in order to highlight some individual counts. The tags on this graph, through their position, symbolise the approximate value of the web count for selected places. Bern, being the capital of Switzerland and an important tourist destination has a place on the top. The

two bar charts in Figure 0.3 show us the resulting web counts from the SwissNames and GeoNames datasets for toponyms. In terms of the web counts, both datasets exhibit very similar characteristics, with few toponyms that yielded zero counts for German and English and many zero valued web counts for Italian.

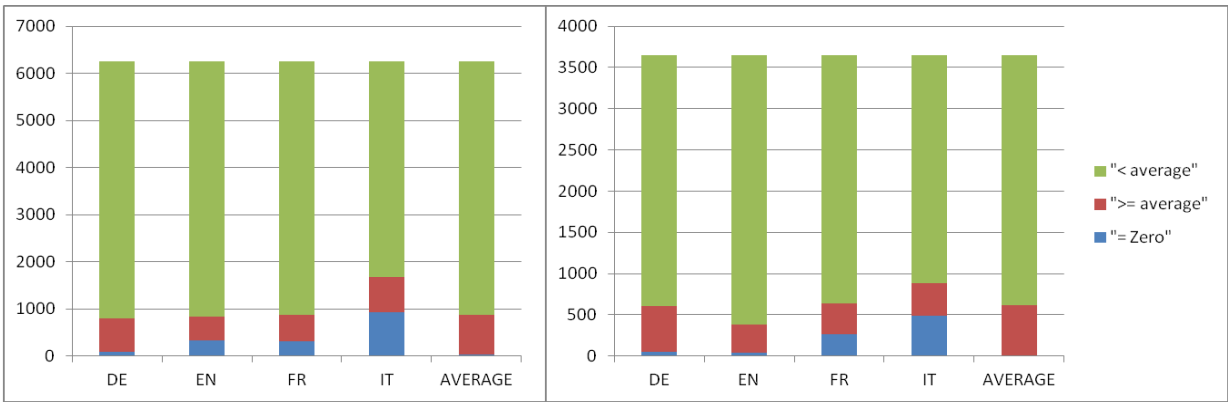


Figure 0.3 - Bar charts showing the web count summary by language for SwissNames (left) and GeoNames (right).

Figure 0.4 and Figure 0.5 show the coverage map of Switzerland, using the web counts generated from the SwissNames and GeoNames dataset, respectively. This is presented with the help of a graduated circle map, where the size of the circle depends on the average web count of all the four languages, for a given toponym, with legend values decided by the Jenks classification method. In both datasets, toponyms that had cumulative web counts equal to 0 were eliminated, but in both datasets this was rare (cf. Figure 0.3). Hence, the lowest value for the average web count was 0.25. There were many toponyms for which the web counts were 0 for three languages and high for the fourth language, which happened to be the language spoken in that area. This is because many tourism-related toponyms are in the local language (i.e. in German, French and Italian, as opposed to English) and some have complicated names. Also, many of the entries in the POI dataset relate to transportation tourist attractions such as 'Luftseilbahn' (German word for cable car), 'Gondelbahn' (German word for gondola lift), 'télésiège' (French word for a chair lift), etc. These names are given in the local language and yielded 0 or very low counts for other languages. A typical example is the 'Felsenegg Luftseilbahn', which is an important tourist attraction near Zurich. Felsenegg is a place near Zurich and 'Luftseilbahn' is German for the 'cable car'. But English, French and Italian web pages do not use the word 'Luftseilbahn', instead they use the corresponding translated word for 'Luftseilbahn' (cable car). Since 'Luftseilbahn Felsenegg' is the

official name, it is never found in English, French or Italian web pages but yields a high count in German.

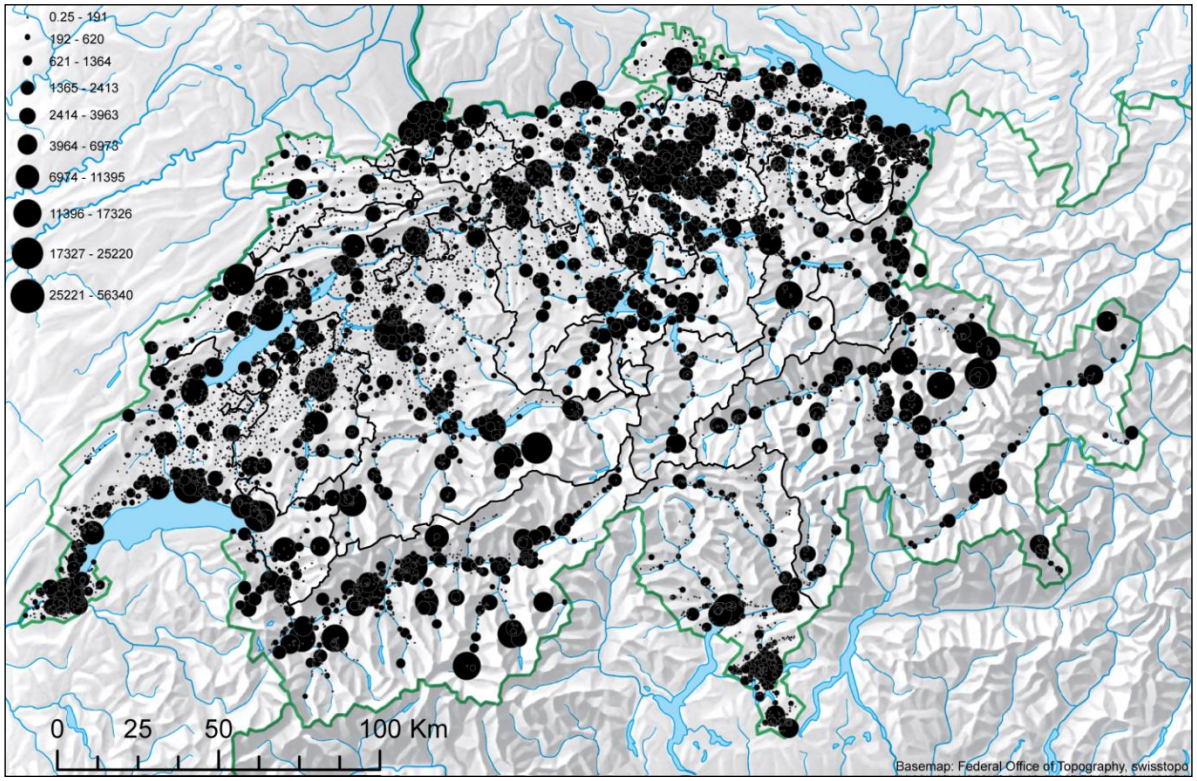


Figure 0.4 - Map showing the geographic coverage by graduated circles of web counts for the toponyms from SwissNames. Size of the circle depends on the average web count of all the four languages, for a given toponym, with legend values decided by the Jenks classification method.

Table 0.9 - List of places in SwissNames and Tele Atlas with the counts of highest frequency.

SwissNames		**Tele Atlas POI**	
Frequency of highest counts	Keywords in different languages	Frequency of highest counts	Keywords in different languages
555	German	290	German
144	French	112	French
16	English	78	English
14	Italian	22	Italian
		185	0 counts
729	Total	787	Total

Table 0.9 shows the number of times a count in a certain language was the highest among the 3 other languages. Out of the subset of records selected, for both datasets the frequency of German web count were the highest and lowest for Italian. There were no web counts that yielded 0 in the SwissNames dataset. Therefore, it can be concluded that the two sets are quite similar in the trends of the web counts presented in Figure 0.3 and Table 0.9, respectively. The number of overlapping toponyms was 2621, which means that more than half of the GeoNames dataset was part of the SwissNames dataset. Hence, for the remainder of the experiment and analyses, only the SwissNames dataset was used, as other datasets seemed to be similar in content or coverage.

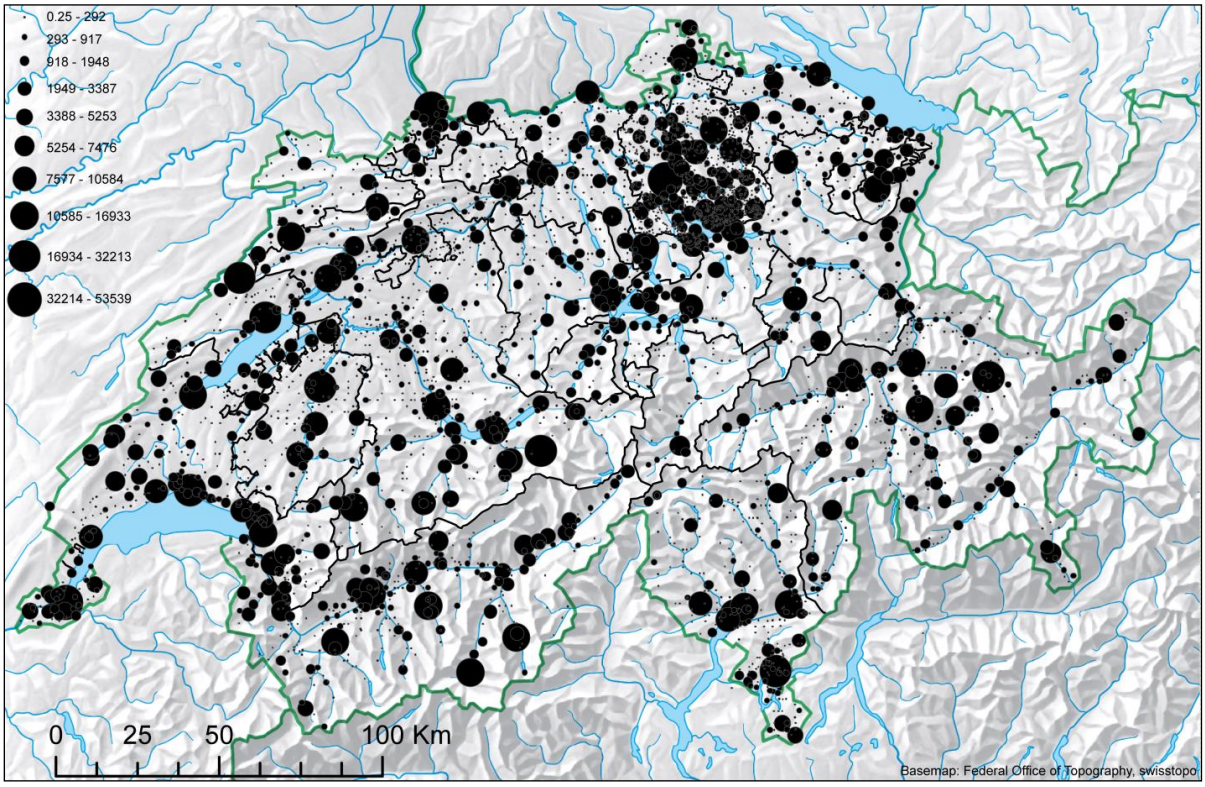


Figure 0.5 - Map showing the geographic coverage by graduated circles of web counts for the toponyms from GeoNames. Size of the circle depends on the average web count of all the four languages, for a given toponym, with legend values decided by the Jenks classification method

Figure 0.3 and Table 0.9 show a comparison between the two datasets. Figure 0.3, attempts to compare the SwissNames and GeoNames datasets. With respect to their toponym content the two lists are quite similar as they have similar 0 values and values that are above and below the average web count. To compare the SwissNames and Tele Atlas POI dataset, toponyms of places with inhabitants more than 2000 people were selected. This resulted in 729 toponyms that were compared with 787 tourism POIs.

1.14 Linguistic coverage

This section, explores Research Question 2 more closely, relating to linguistic differences in web coverage. From the dashed line in Figure 0.1, it can be seen that changes in spelling and translations do not make a difference in the trends. While this is the case for the overall trends, in the case of toponyms with higher web counts, the order changes considerably. Table 0.10 shows the top 10 web counts with toponyms in original names, along with search phrases in different languages. Table 0.11, on the other hand, shows maximum web counts selected from search phrases using the original names, ASCII spelled names and translated names. Toponyms whose counts changed because of the modified spelling, are in bold typeface in Table 0.11. The number of toponyms whose web counts increased, is highest for Italian and lowest for German. This may be because the German speaking region of Switzerland is comparatively the largest and has more toponyms than the other language regions, therefore the probability of a toponym occurring in the German speaking region is high. In turn, many of Switzerland's important places in terms of tourism and population are situated in the German speaking region. On the other hand, the Italian speaking region is the smallest and thus has less toponyms. For 10 toponyms in Table 0.10, 5 of them are pushed down the list when translated into Italian (Table 0.11). Also the toponyms whose counts increased for Italian are not places located in the Italian speaking regions, but are important populated places in Switzerland, such as Zurich. The above observations suggest that web content seems to have toponyms translated into the language being used in the web page, rather than using the toponym in its local language. For example, Geneva when spelt as Genève yields an English web count of 24394 and this drastically changes to 185819 when Geneva is used.

In the map (Figure 0.6) a clear bias of language to region can be seen through web counts that cluster on the German, French and Italian language regions. English web counts, on the other hand, show a dispersed behaviour with similar coverage across different regions. For example, in the case of the German web coverage, it is biased to the extent that important cities such as Geneva (situated in the French speaking part of Switzerland) and Lugano (situated in the Italian speaking part of Switzerland) are hardly visible.

Table 0.10 - Top 10 web counts in four different languages.

Toponyms and web counts used with **German** phrases		Toponyms and web counts used with **English** phrases		Toponyms and web counts used with **French** phrases		Toponyms and web counts used with **Italian** phrases	
Zürich	112074	Basel	64517	Genève	182006	Lugano	31721
Bern	82090	Bern	53620	Lausanne	65637	Locarno	28749
Basel	72679	Lausanne	46037	La Chaux-de-Fonds	53945	Bellinzona	22622
Freiburg	50344	Grindelwald	42305	Yverdon-les-Bains	38541	Chiasso	13736
Luzern	49000	Zürich	33534	Neuchâtel	38084	Mendrisio	13319
Grindelwald	43232	Davos	28232	Montreux	34444	St. Moritz	11961
Glarus	42463	Locarno	26352	Sion	33327	Zermatt	11825
St. Gallen	41291	Sion	26213	Morges	25242	Zürich	10966
Aarau	40025	Lugano	25870	Davos	23031	Ascona	9903
La Chaux-de-Fonds	35223	Genève	24394	Zürich	20051	Basel	9353

Web counts in the French language also show similar behaviour. To better examine this language bias with the language regions kernel density estimates were generated, visualising the χ values (Figure 0.7), calculated by comparing the observed web counts with the average web counts (i.e. the expected number), where the observed web count was the actual web count for any of the languages, and the expected web count was the average of web count, again in each of the four languages. This was done to study the differences between the web counts in the four languages, χ values were computed as follows:

$$\chi = \frac{(obs - \exp)}{\sqrt{\exp}}$$

Table 0.11 - Top 10 web counts with toponyms showing maximum web counts selected from search phrases using the original names, ASCII spelled names and translated names. Toponyms whose counts changed because of the modified spelling, are in bold typeface

Changes for toponyms and web counts with **German** phrases		Changes for toponyms and web counts with **English** phrases		Changes for toponyms and web counts with **French** phrases		Changes for toponyms and web counts with **Italian** phrases	
Zürich	112074	**Geneva**	185819	Genève	182006	Lugano	31721
Bern	82090	**Zurich**	123715	Lausanne	65637	Locarno	28749
Basel	72679	Basel	64517	La Chaux-de-Fonds	53945	**Ginevra**	26650
Freiburg	50344	Bern	53620	**Berne**	49938	**Zurigo**	26544
Luzern	49000	Lausanne	46037	**Zurich**	46146	Bellinzona	22622
Genf	44108	Grindelwald	42305	**Fribourg**	43603	**Berna**	16051
Grindelwald	43232	**Neuchatel**	31712	**Bâle**	38575	**Losanna**	15528
Glarus	42463	Davos	28232	Yverdon-les-Bains	38541	**Basilea**	14764
St. Gallen	41291	Locarno	26352	Neuchâtel	38084	Chiasso	13736
Aarau	40025	Sion	26213	Montreux	34444	Mendrisio	13319

The blue shaded areas denote negative χ values, which in turn means that the expected value was higher than the observed value, while the red shaded areas denote positive χ values, meaning that the expected value was lower than the observed value. For French and Italian web counts the red shaded areas coincide with the language regions. For German, a similar, though less pronounced pattern can be seen. English, on the other hand, is not only almost uniform throughout Switzerland, but also seems to show lesser contrast in the χ values, with values around 1-2 for most of Switzerland. This means that the coverage for English is more evenly distributed as compared to the other languages.

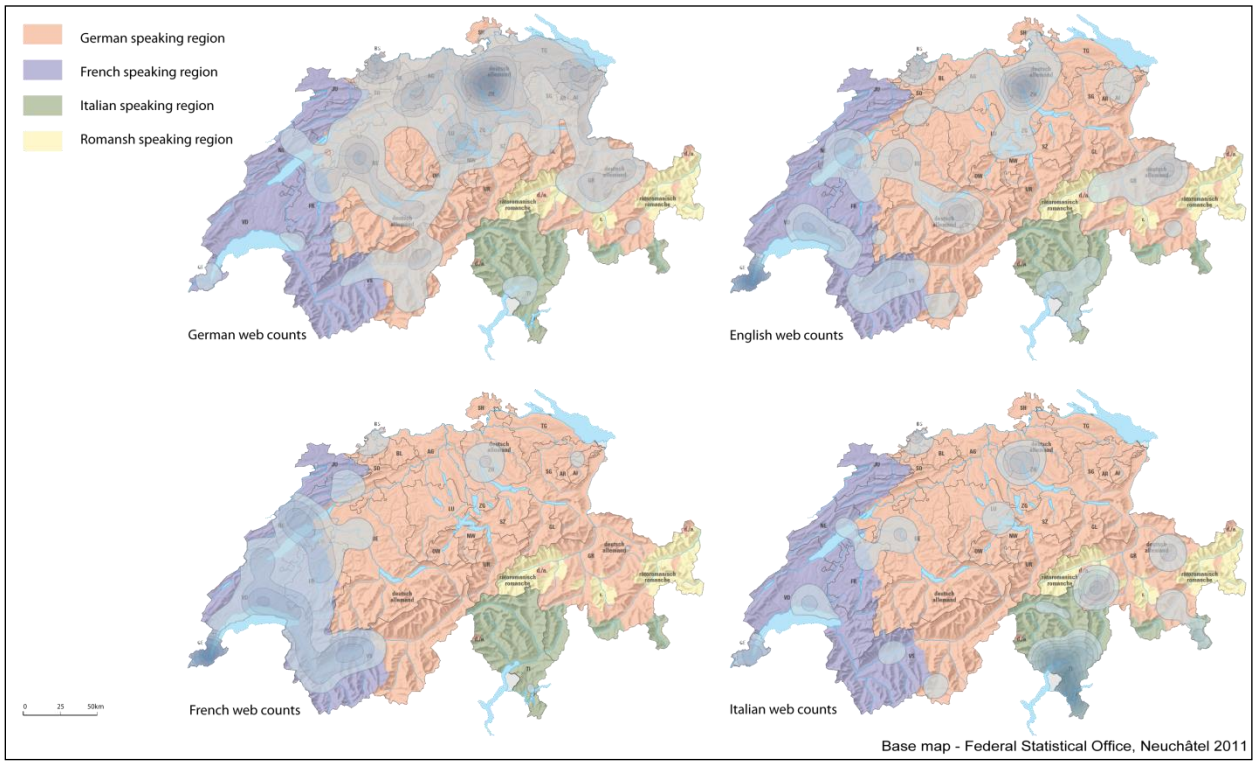


Figure 0.6 - Kernel density map using web counts from GeoNames dataset shown on language region, after toponym spellings were changed

The spatial autocorrelation is the correlation of a web count with itself through space. To measure the spatial autocorrelation, the Moran's I (Table 0.12) was computed. Moran's I always ranges from -1 to 1 and a value near +1 indicates clustering, while a value near -1 indicates dispersion in the values of a variable. To test for the null hypothesis (no spatial autocorrelation), a Z-score was calculated. A Z-score between 1.96 and −1.96 indicates no statistical significance. Looking at the first part of Table 0.12 once can notice that all the points show a clustered pattern except for English. Since the language areas for Italian and French are smaller their Z-scores are very high. To study the spatial autocorrelation in the individual language regions, three sets of points were extracted, by intersecting the toponym points with each of the three language regions. That is, one point set was generated for the German speaking region, a second point set in the French speaking region and a third set in the Italian speaking area of Switzerland. All points except Italian resulted in a high Z-score and no pattern of dispersed points.

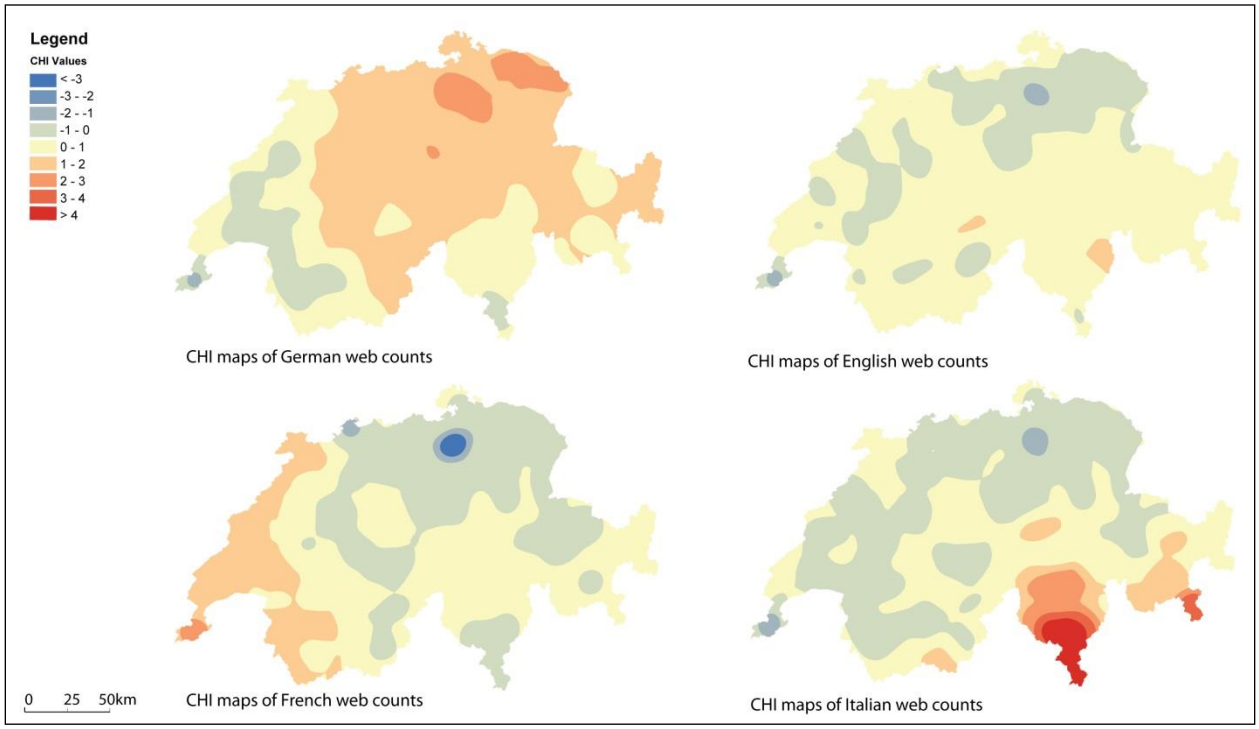


Figure 0.7 - Map showing χ values comparing kernel densities of average and language web counts. Positive χ values (red colours) denote areas where language web counts were higher than average web counts, while negative χ values (blue colours) denote areas where language web counts were lower than average web counts

Table 0.12 - Spatial autocorrelation of language with place (clustered patterns in bold).

	Computed with all points		**Computed with points only in the corresponding language region**	
Measure	**Moran's Index**	**Z-score**	**Moran's Index**	**Z-score**
Average	0.005972	2.402686	-	-
German	0.0020140	7.954030	0.011856	5.4978
English	0.003957	1.615825	-	-
French	0.022078	9.983075	0.009924	2.095638
Italian	0.034569	14.236829	-0.002434	-0.074634

1.15 Influencing factors

This section is devoted to Research Question 3, and thus towards establishing the correlation of web coverage with independent variables. It starts with an analysis of clusters in the web counts

data, in order to gain a better impression of the geographic distribution of web coverage. While Moran's I can give an impression of the global degree of concentration and spatial autocorrelation in a spatial variable, it does not allow to reveal local patterns of spatial autocorrelation. Therefore a measure of local spatial autocorrelation was used, i.e. the Getis-Ord Gi* statistic (Ord and Getis, 1995) on the average web counts across all languages. The output of the Gi* statistic is a Z-score for each point, representing the statistical significance of clustering for a specified distance. Highly positive values denote so-called hot spots, while clusters of highly negative values are termed cold spots.

In the maps of Figure 0.8 and Figure 0.9 one can see that there are a several hot spots, but no cold spots. The hot spots correspond to places such as Zurich, Basel, Bern, Geneva, La Chaux-de-Fonds, Lausanne, Grindelwald, Zermatt, Davos and Lucerne, in effect the top ten counts when all four languages are considered. From the kernel density estimation (Figure 0.6 and Figure 0.7), it was clear that there is a bias towards big cities, irrespective of the language. One potential reason may be that cities have higher populations, hence becoming centres for hotels, transport and offering services related to tourism. To examine this, a comparison between the population statistics and the kernel density diagram of toponym web counts in all the languages, was made. The kernel density estimation was plotted for a 20km radius at a resolution of 1km. Figure 0.7 shows the KDE of average web counts, overlaid on the population distribution of Switzerland calculated for 2003. The most obvious effect is the strong relation between low population densities (< 50 residents/km2) and low values of web coverage. However, there are a few cases where the population is sparse but web counts are high. For example, Zermatt and Appenzell have low population densities, but are very well-known tourist destinations, suggesting that the significance of a place as a tourism destination is also an important factor in determining coverage.

Table - 0.13 reports on the correlation of population and web counts. The second column of Table - 0.13 shows the correlation between the populations of places with the web counts; the third column then gives the correlation between the populations of per canton with the web counts per canton. As a proxy for how significant a tourist destination a place is, the number of rented hotel nights per year for that canton, was utilised. However, it is possible that many of these rented hotel nights were used for business purposes but current works follows the definition of tourists given by the United Nations Conference on International Travel and Tourism, 1963 (Leiper, 1979): tourists are "temporary visitors staying at least twenty-four hours in the country visited and the purpose of whose journey can be classified under one of the following headings: (a) leisure (recreation, holiday, health, study, religion, and sport), (b) business, family, mission, meeting."

The corresponding correlation coefficients are given in the last column of Table - 0.13. Note that data was available only until 2003, hence the contents of Table - 0.13 is for the year of 2003. The statistical data are published by the Swiss Federal Office of Statistics, Neuchâtel.

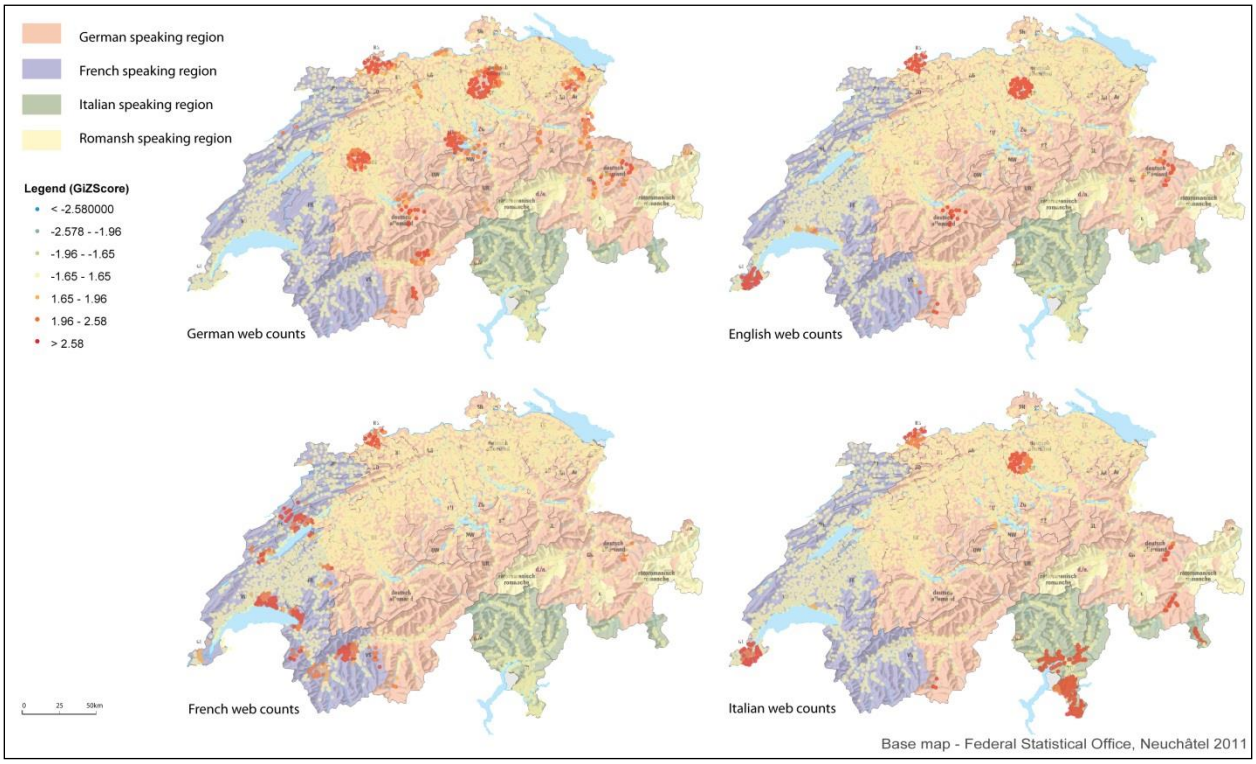


Figure 0.8 - Hotspot analysis of web counts for individual languages. Several hotspots but no cold spots can be seen.

Table - 0.13 Correlation (r) of web counts with population and hotel nights (highest correlation per language highlighted in bold).

Language	Correlation with population (all places with information)	Correlation with population (cantons only)	Correlation with hotel nights rented per year (cantons only)
German	0.3817	**0.6676**	0.4508
English	0.1811	0.2793	**0.5079**
French	0.2056	0.2159	**0.2360**
Italian	0.0612	0.5496	**0.6023**

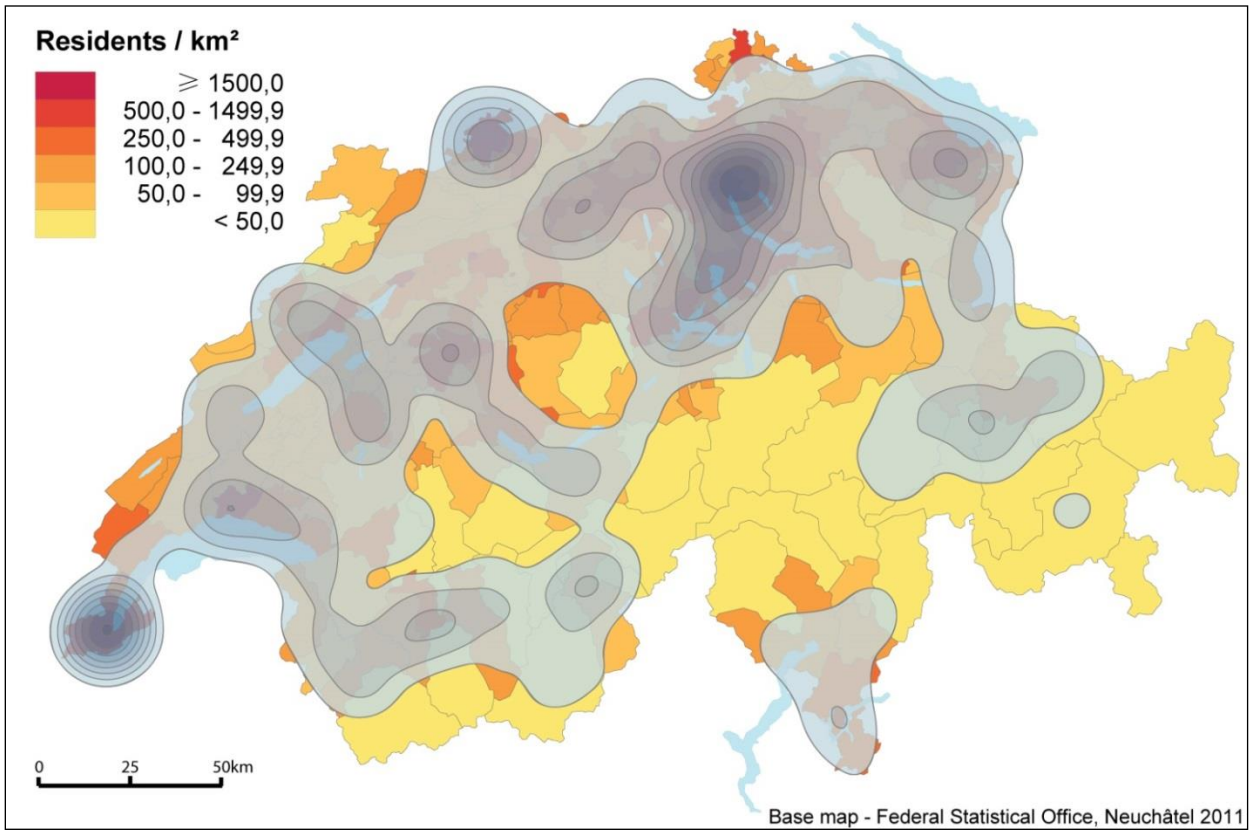


Figure 0.10 - CHI map comparing kernel densities of population and average web counts of Switzerland. Positive χ values (red colours) denote web counts higher than population, while negative χ values (blue colours) denote population higher than web counts. Labelling is approximate

Determining locations and related activities using UGC

1.16 Introduction

Web resources potentially contain interesting information that can be used to enrich spatial databases in order to achieve more personalised portrayal in web mapping or in location-based services (LBS). In particular, traditional geographic information retrieval (GIR) methods (Purves et al., 2007) coupled with UGC (user-generated content) can be used to explore diverse aspects of place. For instance, georeferenced image metadata (e.g. Flickr) can be automatically extracted and annotated (Purves et al.; 2008; Rattenbury et al., 2007) and the information therein can be used to better understand what people perceive as a place and its extent.

In the past, place has been described with the help of geographic elements, qualities and activities (Purves and Edwardes, 2007; Edwardes and Purves, 2007; Purves, 2011). What people do in a place or what a place affords people to perform, is an important aspect of place (Curry, 1996; Jordan et al., 1998) as this is possibly influenced by the infrastructure or some other aspects of the place. There has been little research in linking activities and place (Zhu et al., 2013), and this chapter tries to address this by analysing photographs taken by people to determine what they perceive as tourist-related activities for a given place. The two research questions thus explored are:

1. With the help of UGC, is it possible to extract locations of places and their related activities? Is it possible to assign individual locations to these activities?

2. Having extracted locations and their activities, is it possible to group these locations based on how similar the activities performed there are?

In order to answer these questions we use a methodology that consists of the following steps:

1. Gathering of various forms of UGC for Switzerland for different time periods.

2. Implementation of an automatic location outlining algorithm using UGC, for every toponym from a lookup gazetteer list.

3. Selection of toponym locations based on some measures that ensure minimum accuracy.

4. Selection and grouping of activity terms from various sources.

5. Looking up these activity terms in user generated tags in UGC, thereby associating locations with activity terms.

1.17 Data

1.17.1 Flickr

For the study described in this chapter, georeferenced image metadata, extracted from Flickr and Picasa are considered. Flickr (www.flickr.com) and Picasa (picasaweb.google.com) are picture and video hosting websites with an online community. Users upload pictures in order to store and/or share them with other users. Image metadata, in particular Flickr, has been used extensively in the past for studies relating to place. Image metadata was downloaded using the APIs provided by both websites and metadata information such as the picture id, user id, date of the picture taken, date of the picture posted to website, title of the image and image tags/description were downloaded. For this thesis, Switzerland was selected as the area of study. To obtain the relevant images and metadata for this area, geocoded image metadata was used to restrict the study to images that lie within the bounding box for all of Switzerland viz. NE 47.808380, 10.492030 to SW 45.818020, 5.955870 for the period 2006-2011.

Flickr allows users to tag their images and group them into online sets using these tags. For the following experiments, these user-provided image tags were the essence of the data, as many of these images were tagged with words indicating activities and toponyms. Picasa does not make use of tags but users are allowed to add descriptions in a natural language to their pictures. Tags and image descriptions were used as a means of inferring the image content. No image analysis or image processing was performed in the course of this study. Image metadata for approximately 2 million georeferenced images from Flickr and around 1 million georeferenced images from Picasa was extracted for this study.

Flickr data is known for its participation inequality; i.e. a small number of users provide large numbers of images, thereby dominating the collection, which introduces bias in the data. Purves et al. (2011) report that in their Flickr dataset, 73 percent of the images were uploaded by 10 percent of the users. To remove the bias caused by the most prolific users, they deleted photos uploaded by the top 25 contributors. A slightly different approach was used in the current study.

Users often upload the entire set of georeferenced images from their camera after a trip or event, with the same title and tags for each image. These tags often have many place names and activities in the text because they tend to be tagged together, as a set. As opposed to considering all these images, only the first picture taken on this trip or event was retained, leading to a smaller but de-duplicated set of images. After filtering from a set of approximately 2 million images, half of these were filtered out and approximately 1 million images were used in the experiment.

1.17.2 GeoNames

GeoNames is a geographical database or gazetteer. It was used as a lookup table for toponyms in Switzerland. This gazetteer contains information such as the toponym name, coordinates, feature code, population, canton, and administration.

There are 20,726 toponyms for Switzerland. From this gazetteer, toponyms whose 'feature code' (field in GeoNames), were marked as populated places (PPL), administrative division (PPLA), and capital or political entity (PPLC) were selected for the study. Toponyms with these feature codes covered a majority of tourist points of interest, populated areas, and areas with train stations, and were therefore often tagged in pictures uploaded by users to Flickr and Picasa. The remaining toponyms included names of administrative divisions, building names, and names of peaks among others. These tend to be seldom used by Flickr users to tag pictures. Hence, out of a total 20,726 records, 4,340 were selected for the experiment.

Table 0.1 shows the breakdown of these toponyms.

Table 0.1 - Classification of the toponyms based on their feature codes from the GeoNames gazetteer for Switzerland.

Feature code	Description	Number of records
PPL	populated place – a city, town, village, or other agglomeration of buildings where people live and work	4,313
PPLA PPLC	Administrative division Capital of a political entity	27
Total		4,340

1.18 Initial steps and activity terms

The aim of this research is to explore the relationship of activity to place and not on the method of identification or extraction of activity terms from a text corpus. Research relating to identification of activity terms and extraction from text has been performed extensively in many fields, and has been discussed in the chapter addressing the background (0). In order to answer the research questions discussed in Section 1.16, a list of activity terms is required to match Flickr tags for these terms. A first list of terms came from those that were collected by Purves et al. (2011), through a voting mechanism that categorised terms as elements, qualities and activities. A list of 107 activity terms was taken from their lists. A second list was taken from research work in the field of leisure studies that examine different types of activities people perform during their leisure and a resulting taxonomy, consisting of 12 clusters (Tinsley and Eldredge, 1995). This taxonomy lists 82 activities that were included in the activity list made for the current experiment. Finally, a third list of activities was created because the previous activity lists (lists 1 and 2) missed terms such as skiing, snowboarding, walking, hiking, etc., which, in the context of Switzerland are very typical activities for tourists and locals. Hence these activities were added to include such terms. A list of such terms that are important in the context of Swiss tourism was made by extracting all words ended with 'ing' from all the tags in Flickr, and some web pages from mySwitzerland.com that spoke about tourism in Switzerland, along with their resulting frequencies. As one might imagine, this list had a lot of unsuitable 'ing' ending words, hence, after some manual effort of deleting and de-duplicating, the three activity lists were put together, resulting in a final list of 424 activity terms.

As a preliminary step towards examining the tagging behaviour of Flickr users, activity terms, and their behaviour, Flickr points were extracted within a bounding box (approximate) drawn manually from prior local knowledge of the extent of 11 places in Switzerland for various time periods between 2009 and 2011. For the same time periods Flickr points for the bounding box covering all of Switzerland were also extracted.

The Flickr points extracted for the bounding box of Switzerland needed to be examined. A simple test was performed to this end, to determine the number of wrongly tagged images. For each toponym term of the 11 toponyms for which bounding boxes were drawn manually, a term search on Flickr tags was carried for that toponym. This resulted in 11 sets of Flickr points and for every set, the number of points outside the manually drawn boxes was computed. For all of the 11 toponyms, the number of Flickr points outside the bounding boxes was between 10-15% of the

original number of images. This metric plays a role in decisions made that are discussed in the later parts of this chapter (See Section 1.19.1).

Table 0.2 - Table showing comparison between top 20 activities in British Isles and Switzerland, using initial list of activities from Purves, 2011. Terms in grey boxes denote changes between filtered and unfiltered lists.

Top activities for the British Isles (unfiltered)	Top activities for the British Isles (filtered)	Top activities for Switzerland (unfiltered)	Top activities for Switzerland (filtered)
music	party	Travel	travel
party	music	Hiking	hiking
gig	gig	Music	music
birthday	wedding	Concert	concert
wedding	birthday	Event	event
christmas	travel	Party	holiday
travel	christmas	Festival	festival
concert	concert	Holiday	trip
rock	holiday	vacation	rock
holiday	festival	Trip	vacation
festival	football	Show	sport
football	vacation	Airshow	bike
vacation	livemusic	Sport	party
gigs	rock	Rock	show
club	club	Bike	cycling
show	work	wedding	christmas
livemusic	cycling	football	tour
work	trip	Tour	explore
drinking	sport	Race	climbing
rugby	show	climbing	walking

After extracting these points, the tags for all the images were examined and compared with a list of activities generated in previous research for places in the British Isles (Purves, 2011). The activity terms in the lookup list were compared to the tags on the Flickr images and a frequency

count was taken for every term in the lookup list. Local knowledge was applied to look at activities in particular areas and for the whole of Switzerland. This test was also conducted in order to draw a comparison of the trend in activities between Switzerland and the British Isles. Differences are expected, given that the activity language in the lookup set was English and this study was done in the context of the British Isles, while the set it was compared to was for places in Switzerland.

Table 0.2 shows a comparison between the top 10 activities in the British Isles and Switzerland. The terms shaded in gray show changes in terms due to filtering. The term ‘airshow’ is a typical example of users tagging many pictures based on single events that yield a completely different value after the filtering process. In the filtered list it is ranked 60^{th} as compared to 12^{th} in the unfiltered list. On the other hand, activity terms such as wedding, football and race are also pushed down in the filtered list but these are still popular activities as they are all rank below 30 in the filtered list.

Although the official languages in Switzerland are German, French, Italian and Romansh, Flickr and Picasa metadata content pertaining to Switzerland have a significant share of entries in English. Hence, unlike previous research on web coverage (Venkateswaran et al., 2014) in 0 that was performed for various languages, it was decided to conduct the remaining research using English terms, due to the nature of the metadata and the results in Table 0.2. In this table 11 out of 20 activity terms are common across the activity popularity lists of the British Isles and Switzerland.

1.19 Methods and experiments

In order to answer the first research question on activities and their location, locations of toponyms in the form of bounding boxes were extracted from Flickr and Picasa content. For each such bounding box, the next step was to extract its activities and locations. One could argue that it might be easier and more accurate to make use of administrative boundaries for a given place, but the aim of this study was to examine people's perception of activities and place through UGC. Furthermore, there has been research, discussed in an earlier section (1.7.3), about people's perception of place and its extent varying greatly with its actual location and administrative boundaries. The next section discusses the approach of automatically extracting locations from image tags.

1.19.1 Automatic shrinking bounding box (ASBB) approach

In this study, a method that we call the automatic shrinking bounding box (ASBB) approach determines locations of places by making use of bounding boxes. The inputs for this method are a set of toponyms in Switzerland and an array of georeferenced image metadata, extracted for the bounding box of Switzerland. This is shown in the workflow diagram (See Figure 0.1) and is part of the data selection process. For every toponym read from the GeoNames gazetteer, a set of toponyms was generated by translating the toponym to other languages, or by spelling the toponym in ASCII form without diacritics (See 1.12.2). Although research performed in light of activity terms was conducted in English, toponyms on were taken in all available languages. The toponym entry for ‘Zürich’ in the GeoNames gazetteer consisted of ‘Zurich’ in the ‘asciiname’ field and a set of 50 entries in ‘alternatenames’ field for Zurich translated into various languages. All of these were taken in the set of toponyms that would later on be used to search through image tags. This was done for every toponym in the gazetteer.

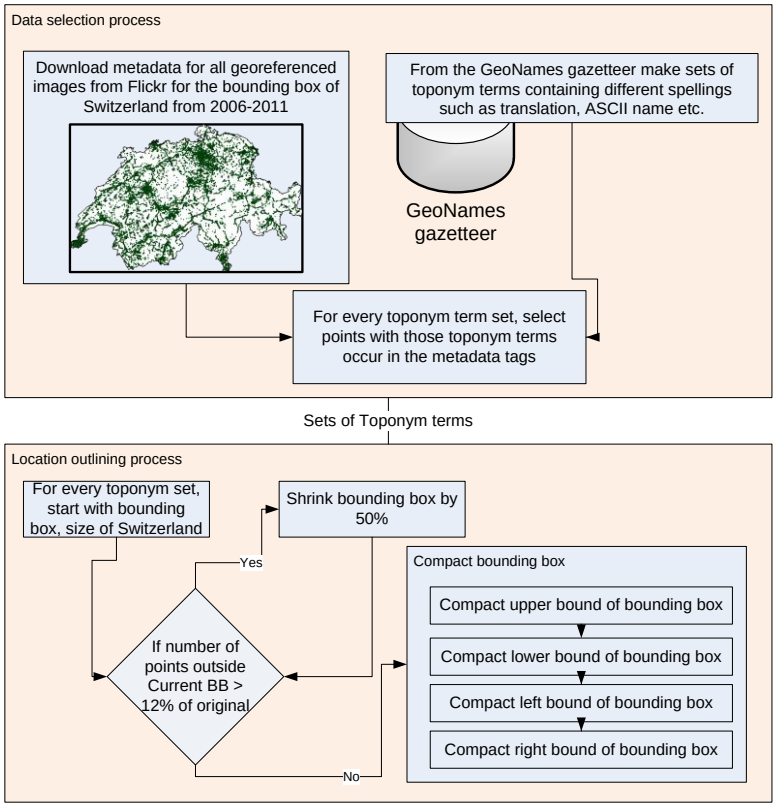


Figure 0.1 - Workflow diagram for the ASBB method, showing the data selection and location outlining process

For every toponym in GeoNames and its multiple spelling variations, only those entries from the images were selected, that had the toponym as part of the tag text in the metadata filtered for outliers. Then a bounding box was drawn around Switzerland and this box was continuously shrunk by 50% per iteration, maintaining the centre, till the best fit around these points was found.

After that a compaction method was used to make this box smaller and find the best fit solution for that particular toponym. This is also shown in the workflow diagram (See Figure 0.1) and is part of the place outlining process. Steps by example of the place outlining process has been illustrated for Zurich in Figure 0.2a.

The stopping criterion for this approach was to have at a maximum of 12% of the points outside the bounding box. This value was determined as part of the initial study discussed in Section 1.18, where, for 11 toponyms whose bounding boxes were manually drawn based on local knowledge, on average, 10-15% of the Flickr points containing those toponyms term lay outside the manually drawn bounding box.

Figure 0.2b shows the boundary of the bounding box along with the administrative boundary of the City of Zurich and Canton of Zurich. In the next section more is discussed on the accuracy of these results, the tolerance of this method, and a mechanism for filtering out inaccurate results.

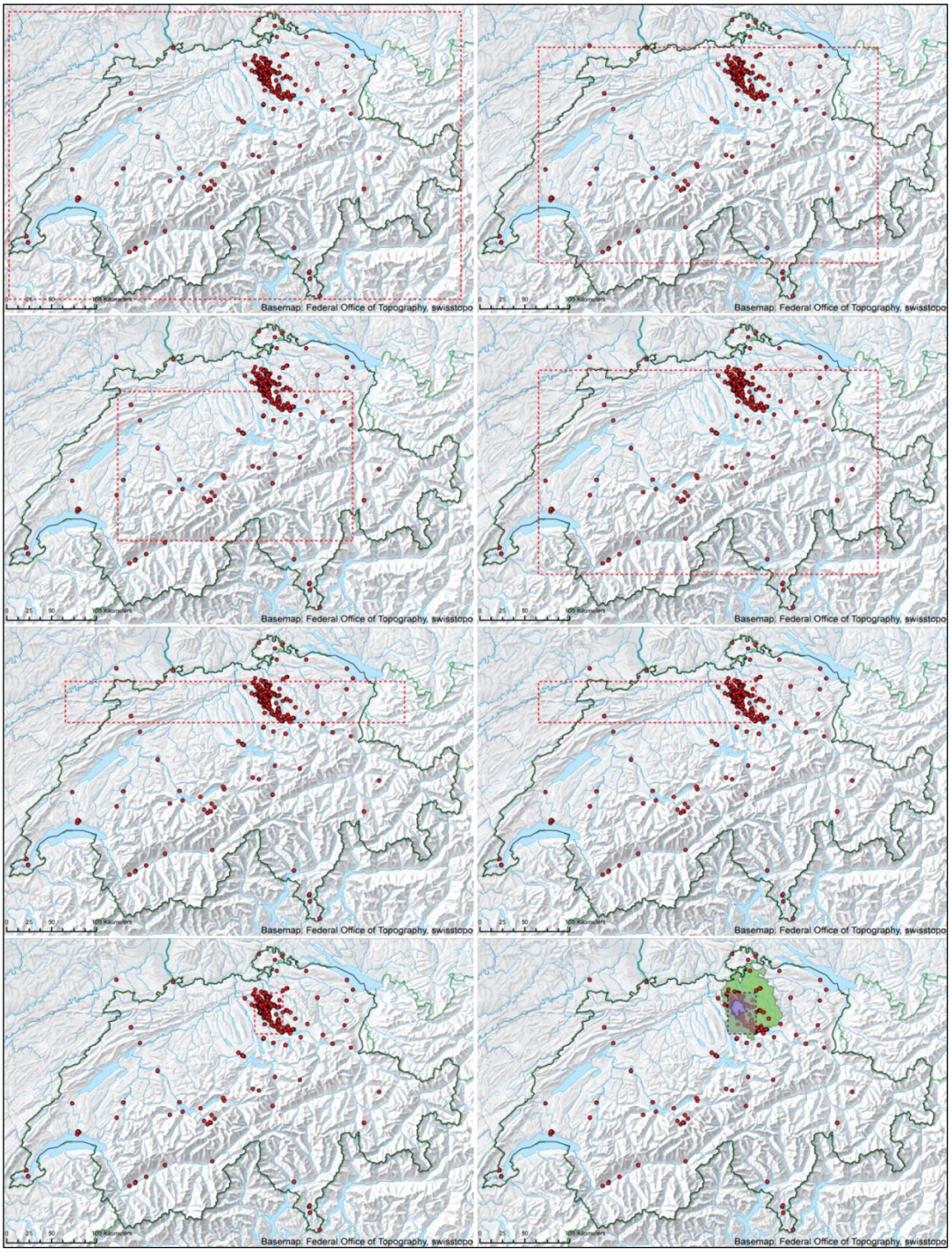
1.19.2 Determining locations from the ASBB method

After the initial selection of places from the GeoNames gazetteer discussed in Section 1.17.2, 4340 toponyms from GeoNames were selected for the experiment. Additionally, another 302 toponyms were deleted due to geo-geo ambiguities, such as duplicate toponym names. Some examples of these are Aarberg, Zell, Wil, Stocken, and Thal. A total of 4038 toponyms were finally used for the ASBB approach to automatically generate bounding boxes. The method generated bounding boxes for 2,148 toponyms. In cases where the method yielded no box for a toponym, one or more of the following reasons was applicable:

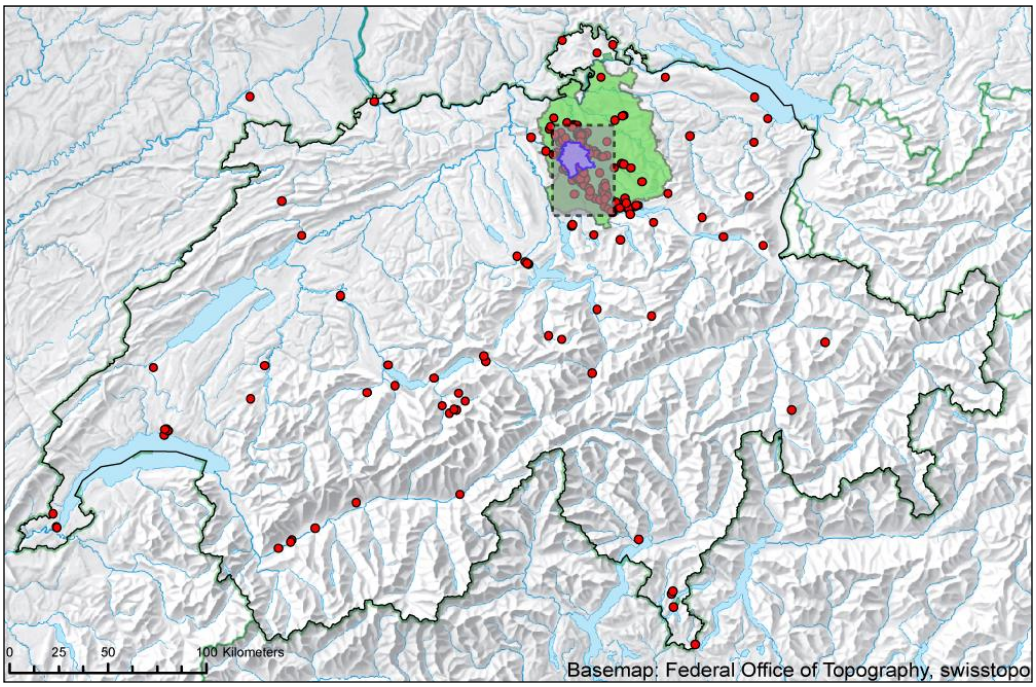
1. If the number of Flickr points containing a set of toponym terms was less than 5, then the experiment was programmed not to proceed.

2. Many of the toponyms in the gazetteer suffer from toponym ambiguity i.e. they were geo-geo ambiguous or non-geo-geo ambiguous. This means that for both ambiguity cases, points containing those toponym terms were scattered all over Switzerland, rather than being concentrated at one point. This yielded a bounding box whose size was similar to

that of Switzerland. The experiment was programmed to not proceed, if the size of the bounding box was bigger than 25% of the area of the bounding box of Switzerland, mainly because no place in Switzerland is that big.



(a)



(b)

Figure 0.2 - (a) Map of Switzerland showing the bounding box (red box) iterations for the example of Zurich using the ASBB method. (b) Map showing the extent of the Canton of Zurich (green) and City of Zurich (purple) and the final bounding box of Zurich obtained through the ASBB method.

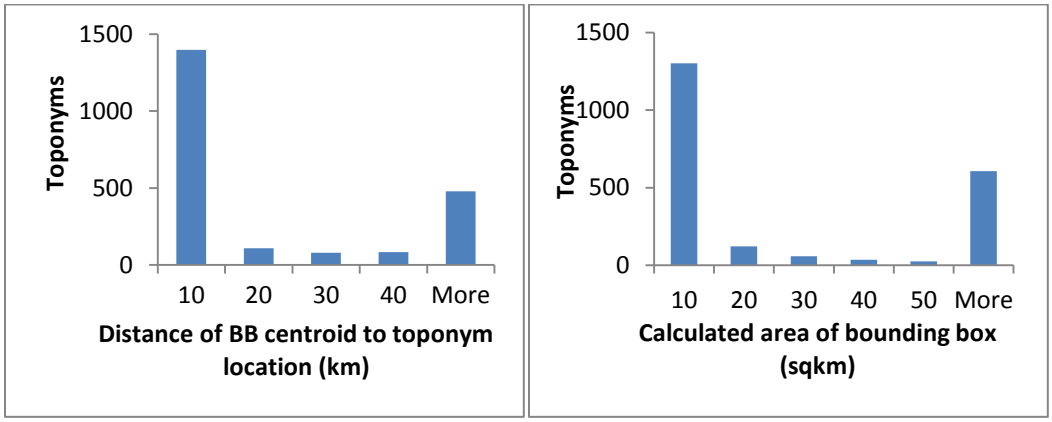


Figure 0.3 - Histograms showing distribution of a) Euclidean distance of bounding box centroid to toponym location and b) Calculated area of bounding boxes

More on toponym ambiguity is discussed in the web coverage chapter (see 1.12.3). For the remaining 2,148 toponyms for which bounding boxes were generated using this method, following analysis was conducted to establish how accurate the method was.

1. Bounding box area

   For every bounding box that was automatically calculated using ASBB, the area of each bounding box was calculated and its distribution is shown in Figure 0.3. For further analysis, only the bounding boxes with area less than 40 sqkm was considered automatically. Bounding boxes larger than 40sqkm were manually examined against the original size of the toponym and not taken if too big compared to the original size. This was done because relatively speaking Switzerland is a small country and cities and towns are comparatively smaller.

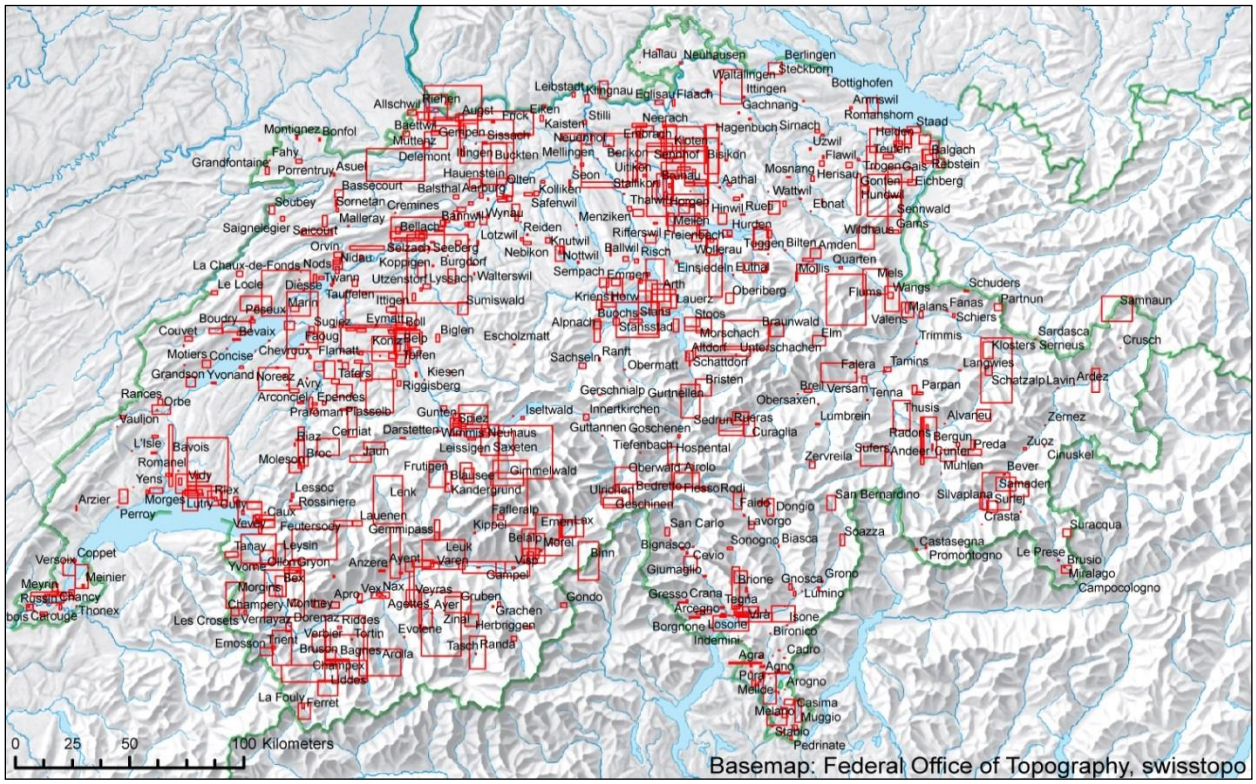


Figure 0.4 - Map of Switzerland showing toponyms and their bounding boxes in red generated using the ASBB method after filtering using the area and centroid filters.

2. Bounding box centroid

   For every bounding box that was automatically calculated using ASBB, the centroid of the bounding box was treated as the location of the toponym. The GeoNames gazetteer has the coordinates of each of those toponyms, from which the Euclidean distance between those coordinates and the bounding box centroid could be calculated. The distribution of this Euclidean distance is shown in Figure 0.3. For further analysis, only those bounding boxes were considered, whose centroids were less than 10km from the actual location. We decided to not consider those above 10km as Switzerland is a relatively small country, with an area of 41,285 km^2 (extent – NE 47.808380, 10.492030 to SW 45.818020, 5.955870). Given the relative proximity of places in a country of this size, centroids further than 10km might be located in another place altogether.

1.19.3 Activities and their locations

As described in Section 1.18, activity terms were extracted using three sources that yielded a list of 424 activity terms. For each bounding box, these terms and their frequency of occurrence in Flickr tags for every point was extracted. The map (Figure 0.5) shows an example of the highest activity frequency occurrence per bounding box. The approximate area of Zurich is circled in red, to show that in the context of activities, the area of Zurich shows typical city like activities like 'film', 'art', 'bbq' and 'food'. The approximate area for the main chain of the Alps[20] in Switzerland is also circled out in blue. The valley in that area is of touristic significance, as many scenic hiking and biking trails pass through it. Table 0.3 shows the top 10 activities for various bounding boxes.

To associate activities with locations, the first step was to determine individual locations of places and commonly performed activities, following which a link could be made between them. To determine locations of places, we devised an automatic shrinking bounding box method (Figure 0.2) as discussed above. One can argue that a bounding box may have not been the best and most accurate way of marking a location. There exist many other ways to extract location from UGC, more advanced and accurate than the current method. Although a bounding box may not have been an ideal geometrical shape to plot for a location (Frontiera et al., 2008), it was relatively straightforward, did not involve a lot of computation and was fast owing to its simplicity. This worked for the current experiment as the aim of the experiment was to study activities in the area, rather than determine the area itself. The entire study is based on the user's perspective, which

[20] http://en.wikipedia.org/wiki/Main_chain_of_the_Alps

makes it seem appropriate to use generated bounding boxes rather than consider the political boundaries of a given place. Figure 0.4 shows the results of the ASBB method. We also calculate the correlation of bounding box areas to the real areas of a location, but did not find a high correlation. We posit that this result is explained by people's perspective of a place not being proportional to the actual administrative boundary of a place.

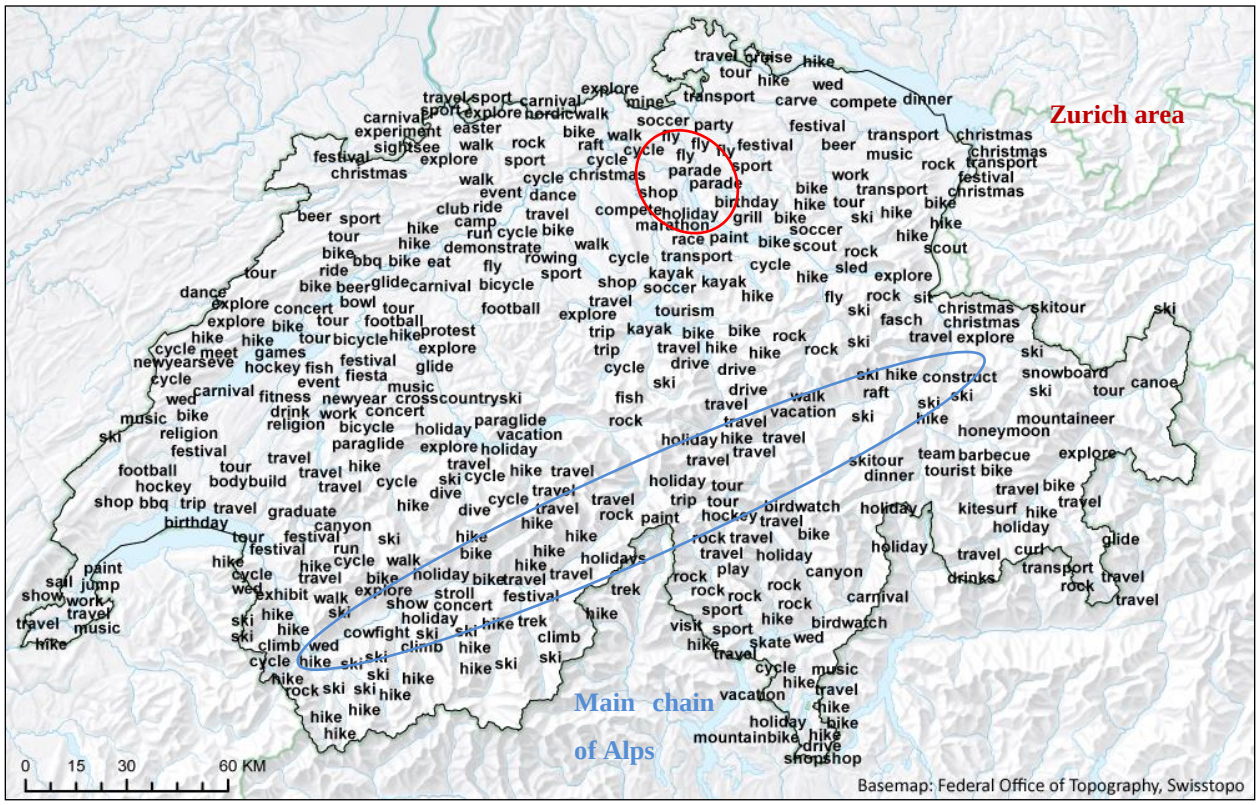


Figure 0.5 - Map of Switzerland showing highest activity per bounding box.

Figure 0.7 to Figure 0.9 shows bounding boxes and their respective activities. The size of the label denotes the popularity of the activity in the particular bounding box. These maps are similar to a visualisation technique adapted to maps by Hahmann and Burghardt (2011). Figure 0.7 shows the bounding box of Zurich, calculated using the ASBB method. The bounding box shows activities typical of cities, with ‘parade’ being the most significant. The Zurich Street Parade is a widely attended techno parade organised in August and is known to be attended by up to one million people. It is also the most attended techno parade in Europe making it a very popular activity. Along with that it is also a typical activity that people photograph a lot, as compared to activities

that require more involvement such as skiing, paragliding etc. that are difficult to photograph by the performer of that activity.

Table 0.3 - Top 10 activities for various places in Switzerland

Zurich	Engelberg	Lucerne	Switzerland
parade	Ski	travel	travel
travel	Hike	trip	hike
music	Travel	kayak	ski
party	Trip	festival	music
christmas	holiday	vacation	concert
bike	snowboard	carnival	event
concert	Sport	transport	holiday
event	Jump	holiday	festival
trip	Rock	paddle	trip
walk	skijump	rock	rock

Figure 0.6 - Bounding box (calculated using the ASBB method) of Les Diablerets showing top 9 most popular activities. Position of labels does not indicate location of activity.

Figure 0.7 - Bounding box (calculated using the ASBB method) of Zurich showing top 20 most popular activities. Position of labels does not indicate location of activity.

Figure 0.6 shows the bounding box of Les Diablerets. Les Diablerets is a village and a ski resort in close proximity to the snow park Glacier 3000. Glacier 3000 is known for several cable cars and ski lifts, ski paths, hiking paths, and alpine coasters. It is also known for associated activities such as skiing, cross-country skiing, biking, hiking, nordic walking, dog sled riding, glacier walking, and glacier flying among others. There is also an annual event known as the Glacier 3000 Run that is a popular mountain marathon/running event.

Figure 0.8 shows the city of Lausanne and its surrounding area. The bounding box of Lausanne is marked through the extent of the background map. Although the centroid of the bounding box is spot on, it is an example of an inaccurate bounding box. The administrative area of the city of Lausanne can be seen marked as a beige polygon, and it is clearly much smaller than the bounding box. However, in the context of activities, the terms are quite close to describing Lausanne. Lausanne is very well known for concerts, theatre and music festivals. Most of the other activities are, once again, typical of cities.

Similarly Figure 0.9 shows the bounding box of Lucerne. This is similar to the case of Zurich where the canton and city share the same name. Therefore, the bounding box of Lucerne is much bigger than the city of Lucerne as points are spread across the whole canton.

Figure 0.8 - Lausanne and its surrounding area, showing top 20 most popular activities. Bounding box of Lausanne (calculated using the ASBB method) is marked through the extent of the background map. Beige coloured polygon shows the administrative area of the city of Lausanne. Position of labels does not indicate location of activity.

But even though these maps with tag clouds overlaid on them tell us something about the nature of the place with respect to activities, the actual location of the activity still remains unknown. Figure 0.5 shows the highest activity per bounding box. This map shows some interesting trends in how the activities change with respect to the terrain, place demographics, proximity to water bodies, and other such factors. Figure 0.5 only gives an approximate idea of the locations of activities and generates only one activity per bounding box.

To further examine the activities for a particular bounding box, we studied activities for the bounding box of Zurich. For the calculated bounding box of Zurich, all Flickr points with tags that contained the term ‘parade’, ‘shop’ and ‘hike’ were extracted. These activity terms were selected as they are very different from each other. Using these points, a kernel density estimate was plotted with the aim of outlining possible locations where these activities take place. KDE (O'Sullivan and Unwin, 2003) is a useful method for observing patterns in point data by pointing out the overall areal density. For our study we selected the bandwidth (also known as r or smoothing parameter) to be 500m as Zurich is a small city (87.88 km2) and most points of interest are in relative close proximity. The map was plotted at a resolution of 10m.

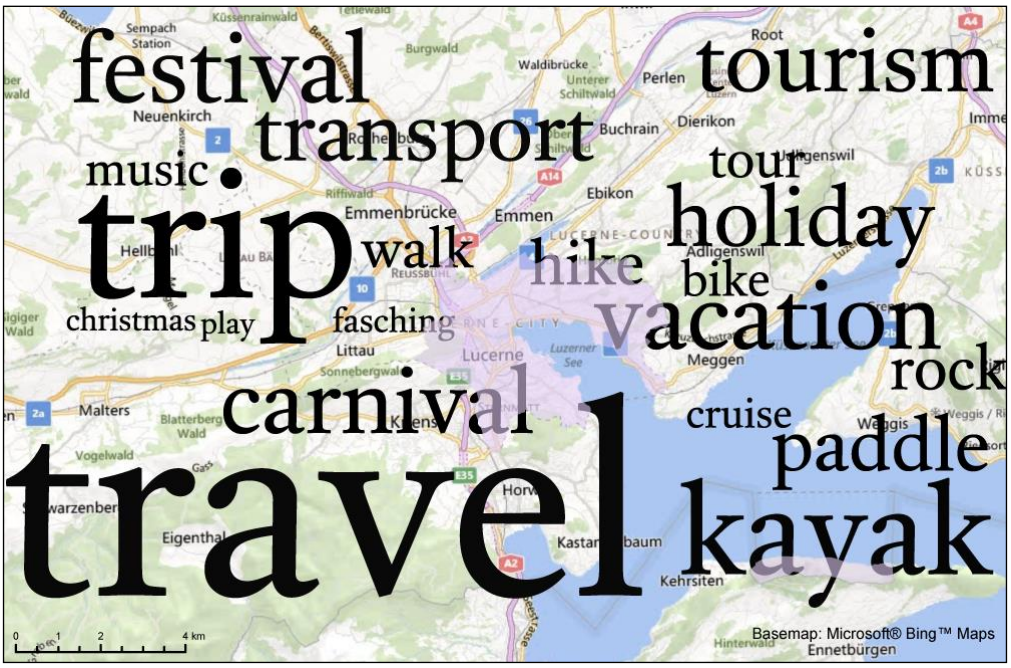


Figure 0.9 - Bounding box (calculated using the ASBB method) of Lucerne showing top 20 most popular activities. The polygon in light pink shows the administrative area of the city Lucerne. Position of labels does not indicate location of activity.

The maps (**Error! Not a valid bookmark self-reference.**) show this result. The first map is a plot of the KDE for all pictures containing ‘Zurich’, ‘Zürich’, ‘Zuerich’ and ‘Zurigo’. This gives us a basic idea of the extent of Zurich in the minds of Flickr users by looking at the density.

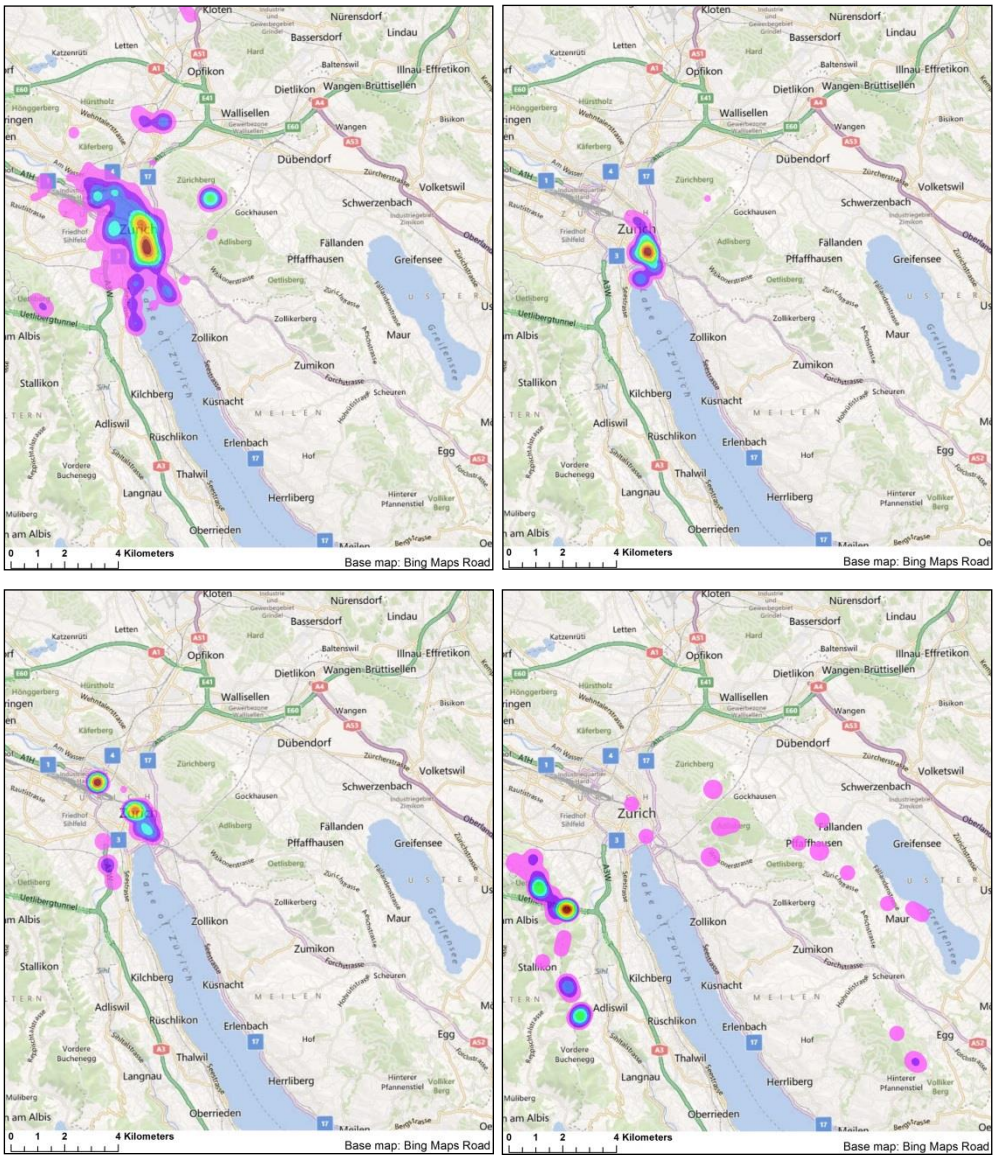


Figure 0.10 - (From left to right and up to down) Bounding box extent of Zurich showing KDE of Flickr tags containing terms related to 1)Zurich 2) Parade 3) Shopping and 4) Hiking.

The second and third diagrams point to the areas which users of Flickr perceive as being places to shop and watch a parade, respectively. Both these maps point to the centre of the city of Zurich, which is indeed where the parade passes through. As discussed earlier the annual techno parade passes through the city of Zurich, is known to be attended by up to one million people, and is therefore widely photographed. In the case of the Shopping KDE, the epicentre of the shopping

map points to the main railway station of Zurich, which is home to a big shopping centre that attracts a number of tourists and locals. In the last map, the area for hiking is outlined. This area is of particular interest, given that the centre of the densest KDE (brown area), points to a place known as Üetliberg. Üetliberg is a POI in Zurich and is part of the Albis mountain range that runs through Zurich. Many hiking trails start and end here and it is the last and highest train station on the Albis. The scattered pink dots are hilly places in Zurich and it is possible to do some hiking and walking in these places. The dot at the southernmost end is another famous hiking point of interest known as Felsenegg. The walking trail from Üetliberg to Felsenegg, also known as Planetenweg, is famous and popular. Felsenegg is also well known as it is the only cable car station in Zurich.

1.19.4 Grouping activity locations

In the introduction chapter of this thesis, we introduced a data enrichment structure that we called the FactsDB (read as "facts database"). The FactsDB was proposed to be an auxiliary dataset, which in turn contains information items (called "facts") about different places linked to a spatial database (such as a Multiple Representation Database, or MRDB) via a gazetteer. This way, the spatial data of the MRDB can be enriched with additional information that, at a later stage, may inform the generalisation process by providing additional semantics about particular places. In our experiment, we attempt to use the information on activities gathered to compare similarity between places. Once similarity between places is computed based on the activities, in future work it will be possible to inform the generalisation process based on this similarity.

To take this further, we performed a cosine similarity analysis on our data to compute a similarity between two toponyms (Bayardo et al., 2007; Derungs et al., 2011). The formula given below shows the calculation of the cosine similarity between any 2 vectors A and B. Each toponym is treated as a vector with activities of varying frequencies. For n bounding boxes, the activities along with their frequencies were listed, resulting in n vectors, and a similarity analysis was performed which resulted in an n*n matrix of values that ranged from 0 to 1. A value of 0 indicates minimum similarity between two vectors, and 1 indicates two fully similar vectors. From the n*n matrix we list 3 vectors or locations Zurich, Zinal and Lugano in Table 0.4 and the top 10 similar locations in their corresponding vectors.

$$cos(\Theta) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

For the location of Zurich, it is observed that the top 6 locations that are similar to Zurich are very close to the city of Zurich. Most of these locations are in a radius of 15km from the city of Zurich. However, Schaffhausen, Basel, and Fribourg are also large cities in Switzerland, but do not exhibit a high similarity to Zurich. From this, one can conclude that the pattern of activities seem to be influenced by geographic distance and the demographics of a place. This is also similar for Zinal and Lugano. Zinal is located in the Swiss Alps and is well known for skiing and hiking trails. On the same lines the locations that are similar to it are small Swiss towns that house ski resorts to places well known for skiing and other forms of winter sports. None of them are large cities or large towns like in the first column.

The three maps (Figure 0.11, Figure 0.12 and Figure 0.13) are a visualisation of the similarity of Zurich, Zinal and Lugano. The boxes denote the bounding boxes that were calculated using the ASBB approach and the colour of the bounding box denotes the similarity to the location under examination; ie. the deeper the colour, the more similar it is to the location under examination. The maps are interpreted in detail in the discussion section (Chapter 7) of this thesis.

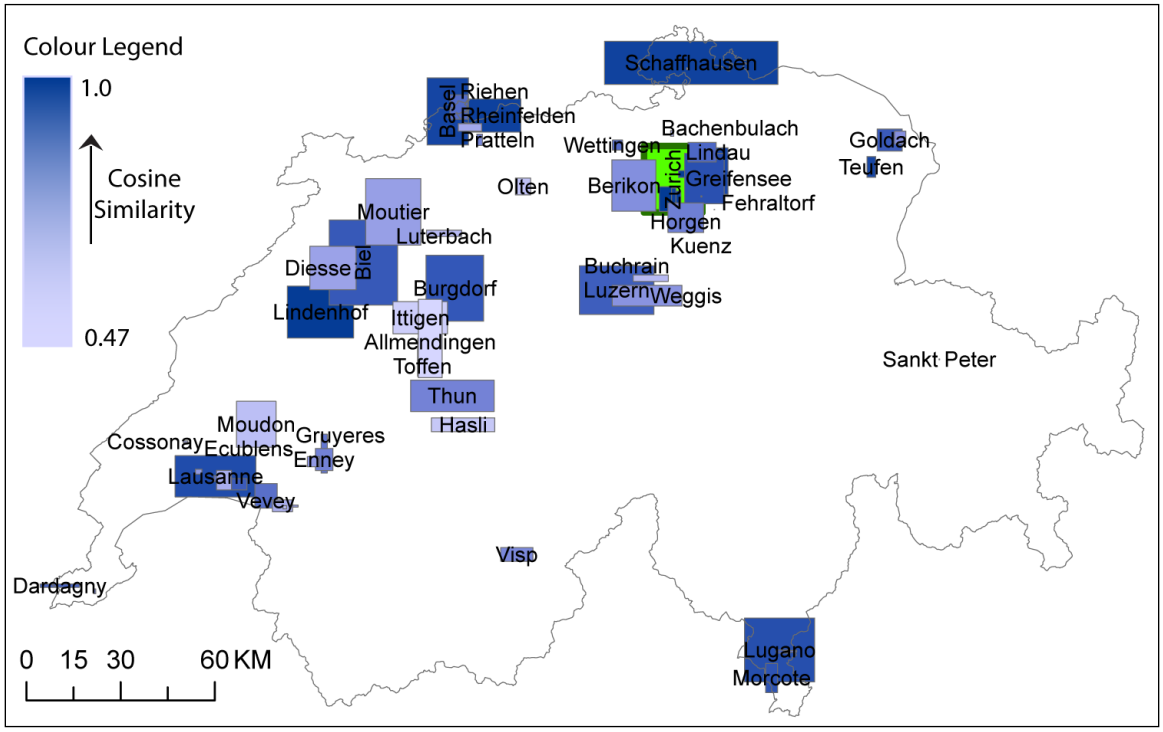


Figure 0.11 - Map of Switzerland showing bounding boxes that are similar to Zurich. Zurich is denoted as a bright green bounding box. The darker the shade of blue, the more similar it is to Zurich.

Table 0.4 - List of locations showing highest similarity to locations in the first row

Zurich	*cos(θ)*	Zinal	*cos(θ)*	Lugano	*cos(θ)*
Adliswil	0.95	Arolla	0.90	Morcote	0.87
Lindenhof	0.89	Flums	0.85	Castagnola	0.86
Gockhausen	0.86	Walenstadt	0.84	Gandria	0.82
Thalwil	0.84	Grimentz	0.84	Fribourg	0.80
Ruschlikon	0.83	Fiesch	0.81	Brissago	0.78
Felsenegg	0.75	Heremence	0.80	Moleson	0.77
Rheinfelden	0.70	Engelberg	0.79	Lungern	0.76
Schaffhausen	0.69	Weissbad	0.78	Biel	0.74
Basel	0.69	Bettmeralp	0.75	Luzern	0.73
Teufen	0.68	Lenk	0.75	Lausanne	0.73

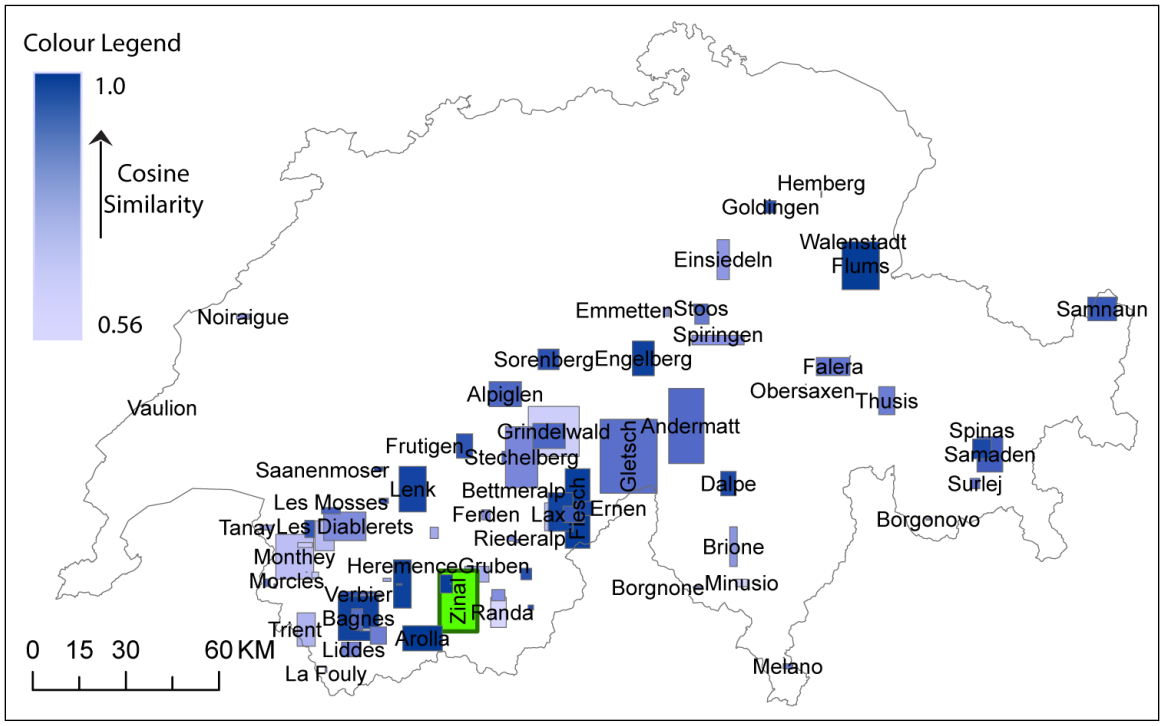


Figure 0.12 - Map of Switzerland showing bounding boxes that are similar to Zinal. Zinal is denoted as a bright green bounding box. The darker the shade of blue, the more similar it is to Zinal.

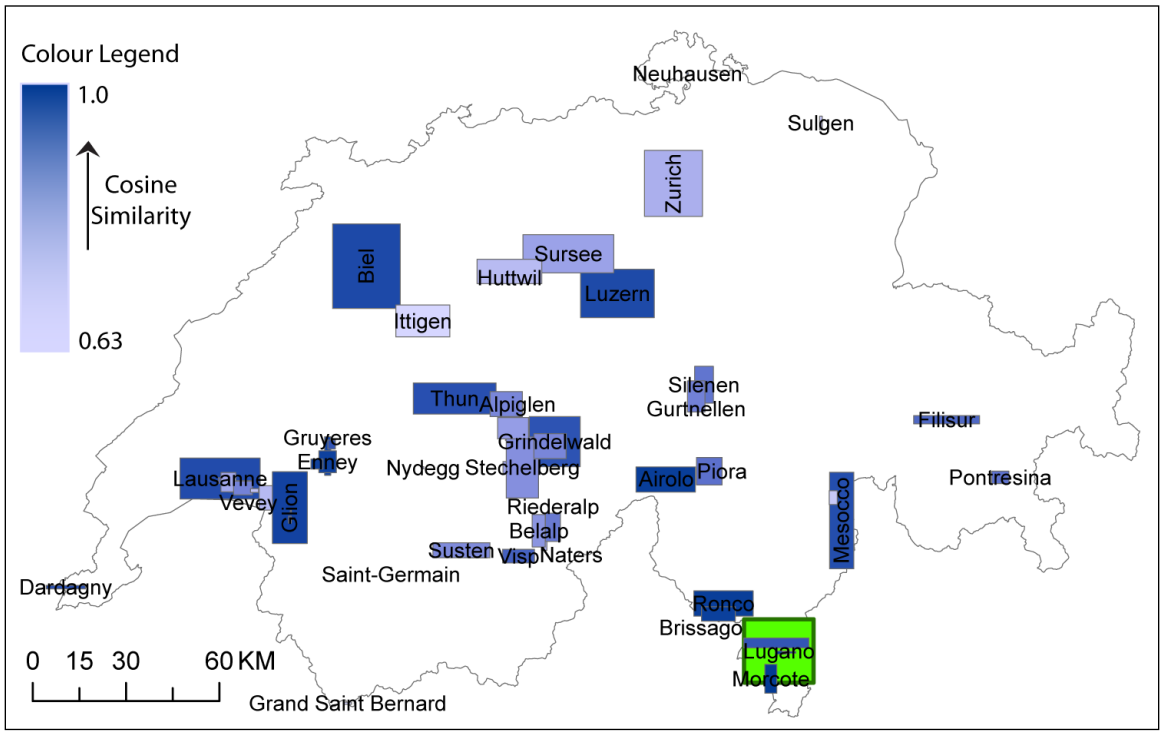


Figure 0.13 - Map of Switzerland showing bounding boxes that are similar to Lugano. Lugano is denoted as a bright green bounding box. The darker the shade of blue, the more similar it is to Lugano

Taking place descriptives further: Exploration of activities

1.20 Introduction

Continuing from the previous chapter, where locations were established from Flickr images and activity terms were then collected per bounding box, this chapter deals more with activities and their interaction with space and each other. Figure 0.1 gives the reader an idea of activities in Switzerland, ranked from the most popular to least.

As discussed earlier, activities are affordances people make about their environment. Hence, they are a way people deal and think of places around them. Having established people's perception of place, this chapter focuses on affordances from activities.

Once extracted, how do these activities behave spatially?

1. How do these tourism-related activities relate to space and how do they relate to topographic data?

2. How can interactions between activities be measured, what is the nature of their interaction and how do they behave spatially and temporally?

In order to answer the above research questions, the remainder of this chapter has been split into four sections. Section 1.21 examines activity infrastructure, particularly for the activity of hiking and similar terms. This section analyses if there exists a meaningful correlation between hiking infrastructure and concentration of hiking related Flickr points.

Section 1.22 focuses on activity terms and analyses the first order co-occurrence between activity terms for different windows i.e. per user, per user per day. Co-occurrence brings out the semantic similarity in terms and we use it to examine interaction between activities. In this section tourist websites are also examined and a comparison between user generated Flickr images and content produced by tourist companies is made.

Section 1.23 draws up a temporal examination of activities, thereby deriving temporal patterns that are also used in the next section.

Finally, Section 1.24 continues from the previous one. Based on the temporal behaviour of these activities, we were able to automatically detect clusters of activities based on their seasons. The last map in this chapter is an example of the motivation discussed earlier on. It is an example of the potential of creating a method of generalisation that focuses on what activity can be performed in a place and not the importance of size of the place per se.

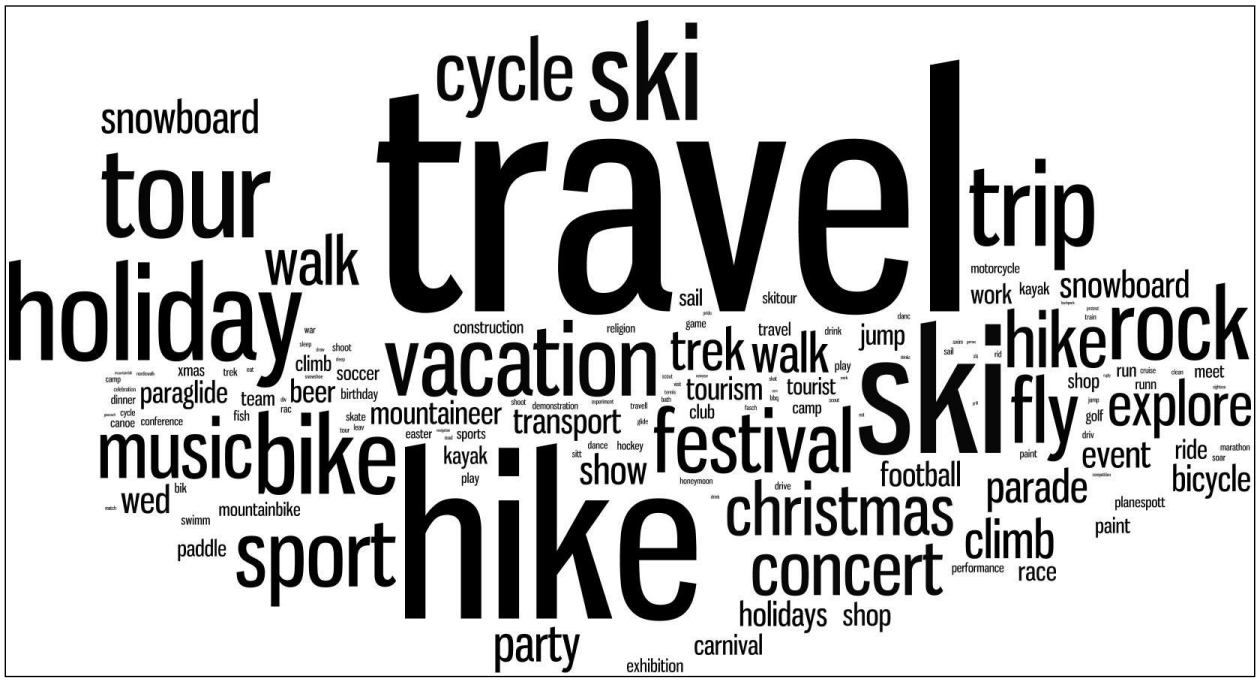


Figure 0.1 - Tag cloud of all activities in Switzerland. Size of activity term corresponds to frequency of occurrence in Flickr tags.

1.21 Infrastructure v/s activity points

The presence of tourist infrastructure could potentially cause a bias in the number of activity points. In the next step we wanted to test if the number of pictures in a certain place is in any way related to the infrastructure provided for a particular activity. We examined the list of the top 10 activities for Switzerland and selected the second highest activity, 'hiking' to analyse this. Using co-occurrence (described below) we were able to make a set of similar activities by selecting those with high co-occurrence values. The other activities that co-occurred highly with hiking were walking, trekking, and mountaineering. Points in Flickr data containing these terms in their tags were extracted. Our aim was to compare the concentration of these points to hiking infrastructure.

VECTOR25[21] contains content and geometry of the topographic map (1:25000) of Switzerland as well as information on man-made and natural features. It is produced by Swisstopo. Road network data in VECTOR25 contains information on more than 70000 km of trails spread across Switzerland. The trails are divided into "hiking trails", "mountain trials", and "alpine trials". This was treated as infrastructure that was available for hiking. To establish if there was any correlation that existed between the number of points and the presence of trails, we divided Switzerland into a grid of squares of size 10km x 10km. For every square the total length of the trails and number of points per were measured, and a correlation was calculated between them. The maps below show the grid, hiking points from Flickr, and trails from VECTOR25, overlaid on the polygon of Switzerland.

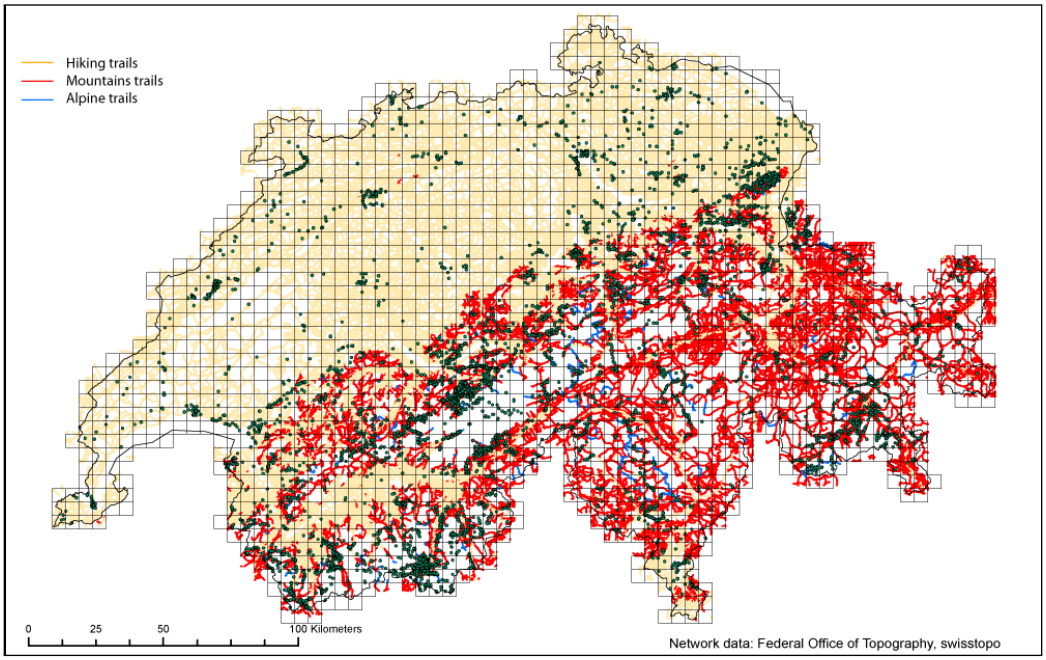


Figure 0.2 - Map of Switzerland showing grid of size 10km by 10km, trails and points from Flickr containing hiking and similar terms. Yellow trails denote walking paths, red trails denote mountain hiking paths and blue trails denote alpine hiking paths.

In all, 32,641 points containing hiking and related terms were extracted from Flickr, and the correlation of trail lengths to the number of images per cell was measured, first for all cells and then for only those cells that had at least one Flickr point. In both cases, although the latter

[21] http://www.swisstopo.admin.ch/internet/swisstopo/en/home/products/landscape/vector25.html

increased, the correlation was low; r = 0.118397 in the first case and r = 0.247039 in the second case. Roughly speaking, the Alps fall in the area where the mountain (red lines) and alpine (blue lines) trails are. We also notice that the number of hiking related Flickr points are high in the Alps region. Hence we performed a second analysis and calculated the correlations again, only retaining Flickr points taken in the Alps and on mountain and alpine trails. The number of points for the second analysis was n = 20,727. The correlation for this case rose to r = 0.51672 in the first case and r = 0.6035 in the second case.

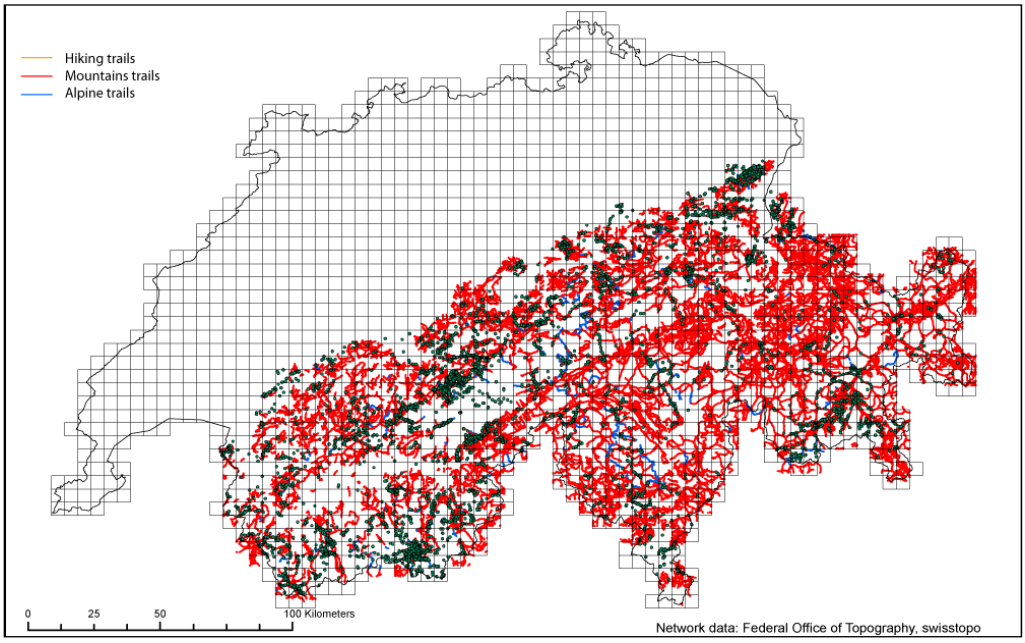


Figure 0.3 - Map of Switzerland showing grid of size 10km by 10km, trails and points from Flickr containing hiking and similar terms only for the Alps. Red trails denote mountain hiking paths and blue trails denote alpine hiking paths.

1.22 Activity interaction through co-occurrence

Co-occurrence can be understood as the frequency of two terms occurring together or in a certain order in a text corpus. While using co-occurrence the window of examination is important. It could be calculated for a single sentence, a paragraph, a page or for the entire file. Determining the window changes the result significantly. In GIR this has been widely used to study terms and their relationship to each other in a text corpus, e.g. in the past toponym disambiguation has been

achieved through the use of co-occurrence (Overell and Rüger, 2007). The main reason for the study of co-occurrence between activities is to get an idea of their semantic similarity. It is safe to say that two terms are semantically similar the higher the co-occurrence values (Spence and Owens, 1990; Lemaire and Denhière, 2006).

In this chapter, the study of term co-occurrence was divided into three parts. In the first part, co-occurrence between two activity terms was calculated for every user. Calculating co-occurrence between two terms is known as a first order co-occurrence. Studies in IR often calculate second and third order co-occurrences, as well as transitive co-occurrences. However, in our research we stick to the simple first order co-occurrence. The second set of co-occurrence values was calculated again between two activity terms, for a window of per user per day. In this case a day was defined as a 24 hour period that lasted from 5.00am on a given day to 5.00am on the next day. 5.00am was selected as the start time as this is around when public transport begins its schedule for the day. Finally, in order to compare the co-occurrence of activities between Flickr (what users photograph and talk about) and what is usually suggested on tourism websites, co-occurrence values for the list of activities was calculated for every webpage from myswitzerland.com. Table 0.1 shows activity pairs for every study, arranged for the top 10 co-occurring pairs.

Table 0.1 - Table showing comparison between top 10 co-occurring activity terms and their values for three sets – Activity co-occurrence per user, per user per day and per record in Flickr and for web pages from a leading tourism website mySwitzerland.com

Per user		Per user per day		Per record		Per webpage	
music:concert	3819	music:concert	2488	music:concert	3510	sport:hike	9121
festival:concert	1752	festival:concert	1329	vacation:travel	1517	walk:hike	4958
music:festival	1657	music:festival	1106	trip:travel	1287	walk:sport	4953
vacation:travel	1557	vacation:travel	1086	festival:concert	1228	sport:ski	4432
trip:travel	1332	trek:hike	1023	music:festival	1181	shop:hike	4082
trek:hike	1148	snowboard:ski	875	trek:hike	1052	hike:event	3937
snowboard:ski	1072	travel:flying	771	travel:holiday	910	ski:hike	3870
travel:holiday	1055	trip:travel	763	snowboard:ski	863	sports:shop	3386
party:music	996	rock:concert	746	travel:flying	813	walk:ski	3322
show:concert	969	travel:holiday	2488	show:concert	3510	tour:hike	3270

To examine the behaviour of activities, activity co-occurrence was studied. The first order co-occurrence between activities for two different time periods is calculated in Table 0.1. In this table there are two observations that are noteworthy. Firstly, the first and second columns denote different time periods i.e. per user and per user per day. Even though the window for the calculation of co-occurrence of terms is completely different the resulting co-occurrences are very similar. From the table it is clear that at least for the top co-occurring pairs, the time period does not matter much, indicating that breaking the information up into per-day counts was probably not needed. Secondly, it can be observed that for activity terms, the co-occurrences of what users photograph and tag in their photos is completely different from what websites suggest as activities to perform. On closer examination it is evident that most of the activities suggested by the tourism websites involve the person's constant physical and mental attention. It may not seem possible to take pictures constantly when one goes skiing or hiking. On the other hand one might take pictures of their favourite rock star or performer or of themselves and friends, while attending a concert of festival. Hence, we posit that the difference in rankings is also because of the nature of the activity.

1.23 Activity behaviour using activity theory

Wang and Cheng (2001) provide basic components of activity theory that have been summarised by Miller (2005). The summary in the form of a table (Table 0.2) has been discussed earlier in 0. Using this table, once notices that time plays an important role while examining activity behaviour. From this table, in the coming parts of this chapter, we examine activity behaviour especially ‘Activity program’ and ‘Activity schedule’.

Temporal patterns

All records that are downloaded from Flickr have multiple attributes such as id, name, date taken, date posted, and tags. We examined the date taken and date posted attributes for temporal patterns. The date taken attribute is the date and time when a picture was taken and it is recorded from the user's camera. Date posted denotes the date and time a particular picture was uploaded to the Flickr website. The graph in Figure 0.4 is a plot of the activity frequency against the month extracted from the Flickr image. In this case the timestamp considered was the image attribute called ‘date taken’.

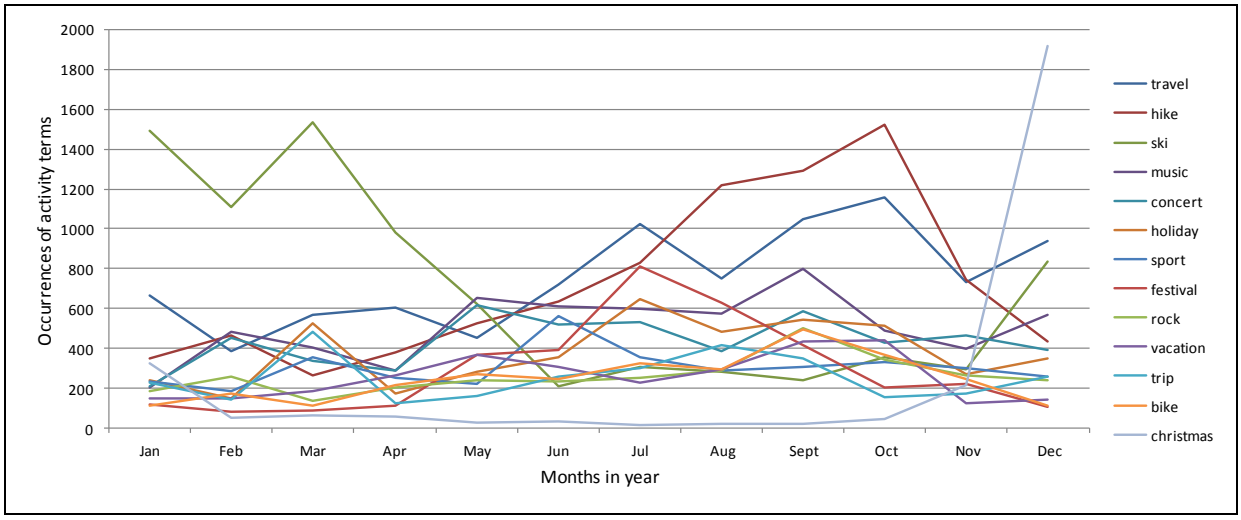


Figure 0.4 - Graph showing plot of activity frequency against the month extracted from the date taken.

The activities in Figure 0.4 show clear differences in behaviour depending on the month. Hiking shows a bias towards the summer and autumn months, but still has some ongoing activity over the year. Christmas shows a clear one-time bias towards December. Skiing shows a somewhat similar behaviour to hiking, except that the bias is towards the winter months. 'Festival' and 'party', on the other hand, seem to be a yearlong activity, with 'party' showing an almost uniform frequency throughout the year.

To complete the examination of the temporal behaviour of activities, it is also important to study the variation of activities during the course of the entire day. Therefore, for a period of one year the activity time for all activities was considered and the results are seen in Figure 0.5 and Figure 0.6. For this study, the 'date taken' attribute of the image under examination is considered. Figure 0.6 shows the trend in activity behaviour during the course of the day. It is no surprise to see that most activities follow a bell curve i.e. rising through the day and the falling off at night. The activity of hiking (green) is an interesting bell curve. There is a sharp increase in this activity during the early morning. This is probably because 5am is when public transport begins operating for the day, when people going on hikes are likely to start their journey. The activity of hiking is highest around mid to late morning and a sharp fall can be seen after lunch time. Therefore, it could be concluded that hikers do most of their hiking activity from morning till early afternoon. Skiing also shows a very similar trend, falling off around 3pm. This is probably because the sun starts to set then and it is probably when skiers stop their skiing activity for the day. The activity of travelling also shows a very similar trend. On the other hand, activities such as 'music' and

'concert' are likely to take place in the evenings and the highest frequency of these activities is between 8-10pm.

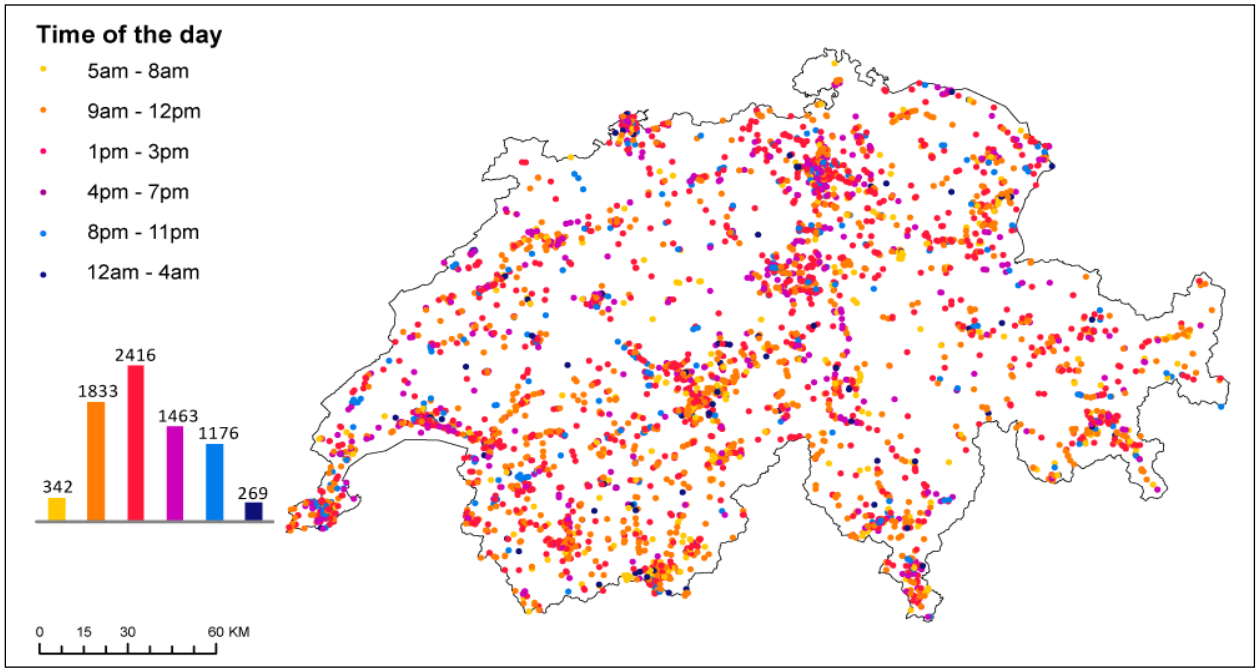


Figure 0.5 - Map of Switzerland showing Flickr activity points. The colour of the dot denotes the time of the day as seen in the legend. The numbers on the bar chart denote the frequency of points at the particular time identified by the colour.

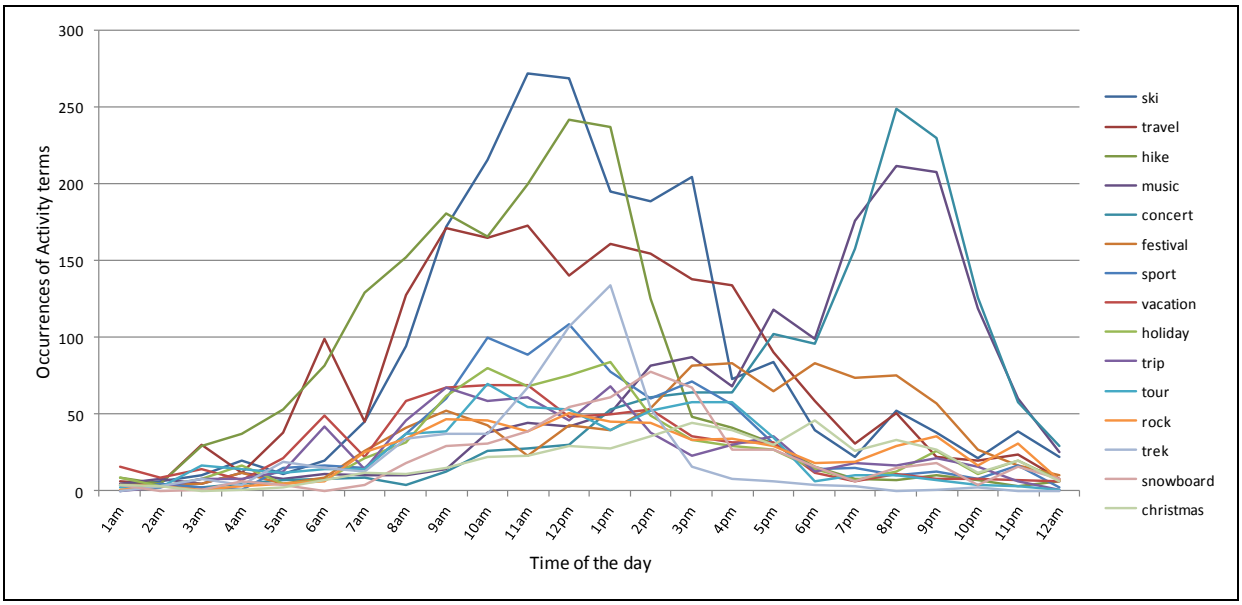


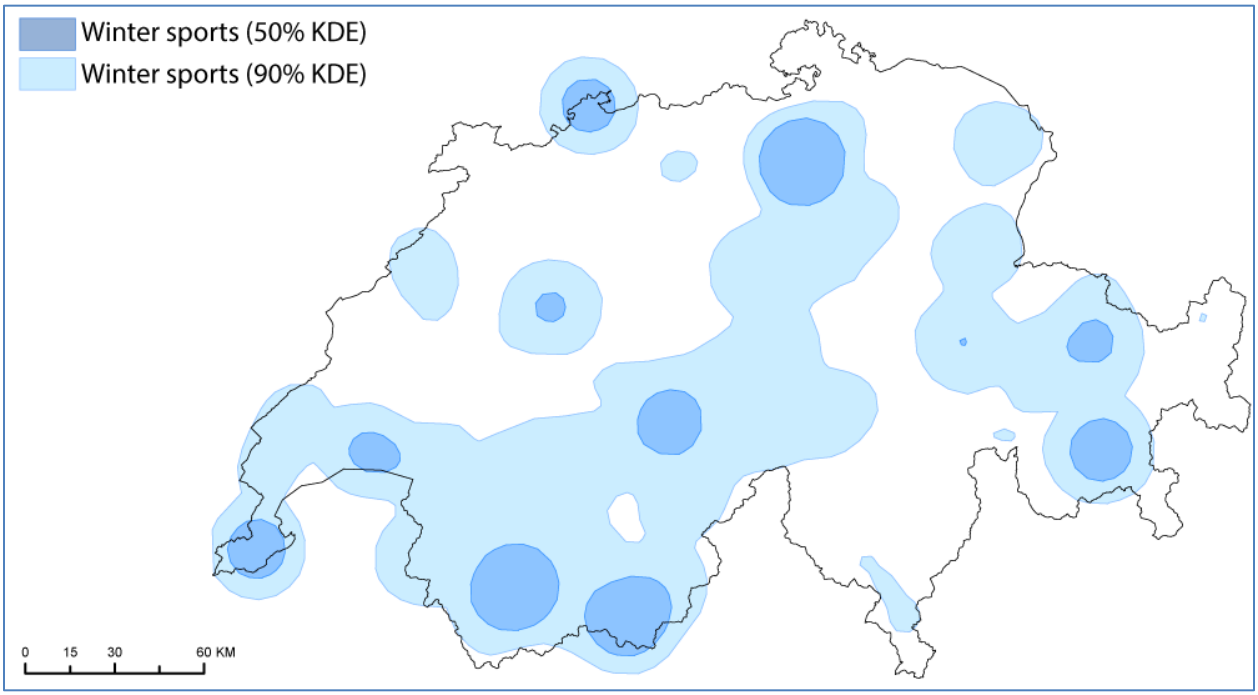
Figure 0.6 - Activity frequency during the course of the day for the year 2010.

1.24 Activity clusters

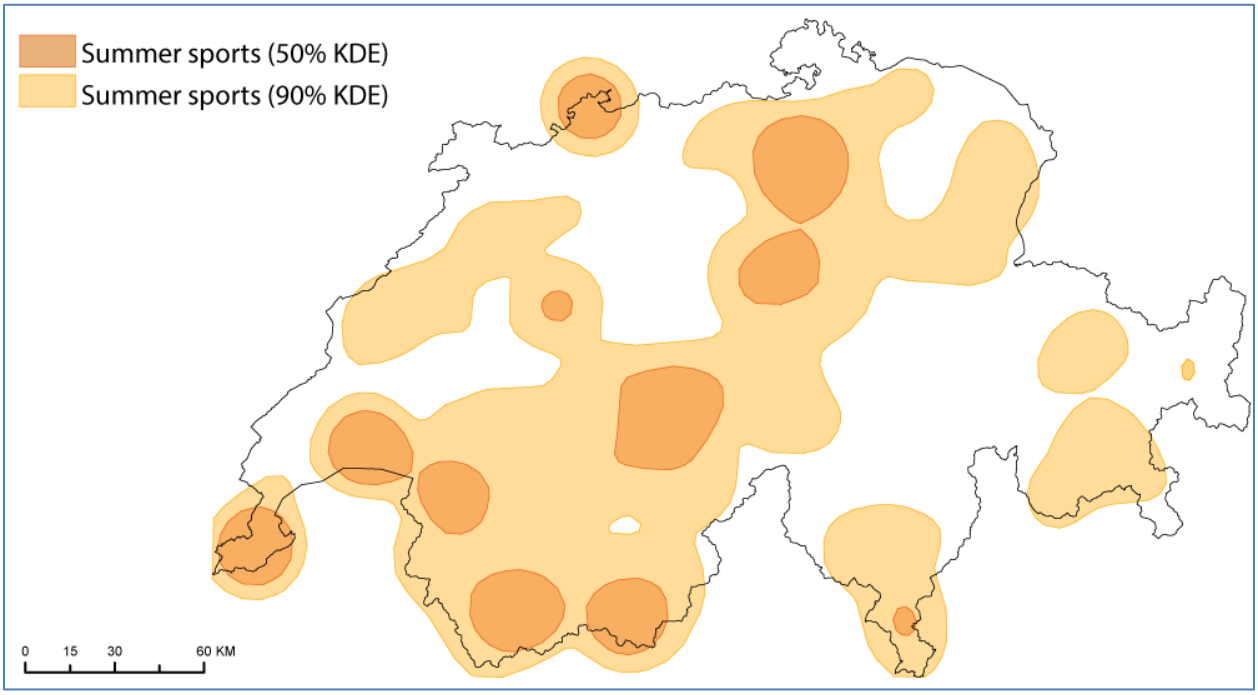
Based on the temporal observations in the previous section (see *Temporal patterns*), a list of summer and winter activities were extracted automatically. Those activity terms, whose occurrences in Flickr tags were highest during the months of June, July, and August were classified as summer activities and those highest during November, December, January, and February were classified as winter activities. The resulting summer activity list had 58 terms and the winter activity list has 36 terms. Some of the examples of summer activity terms are hiking, snorkelling, mountain biking, surfing, rowing, trekking, and climbing and winter activity terms are skiing, snowboarding, cross-country skiing, ski touring, and snowshoe hiking.

However, the drawback of this method of classifying activities is that the timestamp of the individual image comes from the ‘date taken’ and ‘date posted’ parameters. These parameters can only approximately say something about when a picture was taken. For instance, the ‘date taken’ parameter comes from the user's camera and it is possible that it may not have had the right timestamp due to various reasons, thereby resulting in a misleading ‘date taken’ parameter. Similarly it is possible that a user posted their pictures to Flickr, much later than when the picture was actually taken, resulting in a misleading ‘date posted’ parameter. Hence, both these parameters are only approximate measurements of when a picture was taken. For this analysis the ‘date taken’ parameter was considered, under the assumption that most cameras have the right timestamp.

The maps in Figure 0.7 and Figure 0.8 show the 50% and 90% kernel density estimate of the area of sports during summer and winter and the area of water and mountain activities respectively. Particularly 50% and 90% has been chosen as these areas have been often used as an estimate of home ranges and core areas respectively, of animals in behavioural studies (Wartmann et. al, 2014). In this case we have considered the same values to match core areas and home ranges.

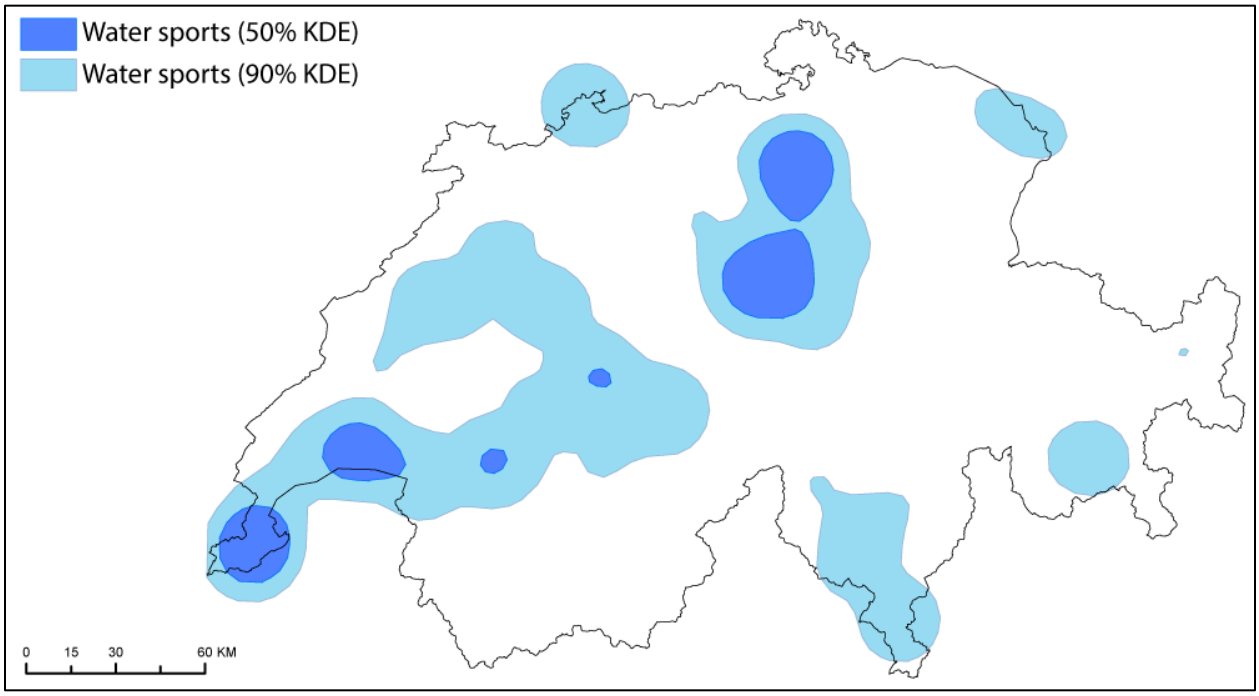


(a)

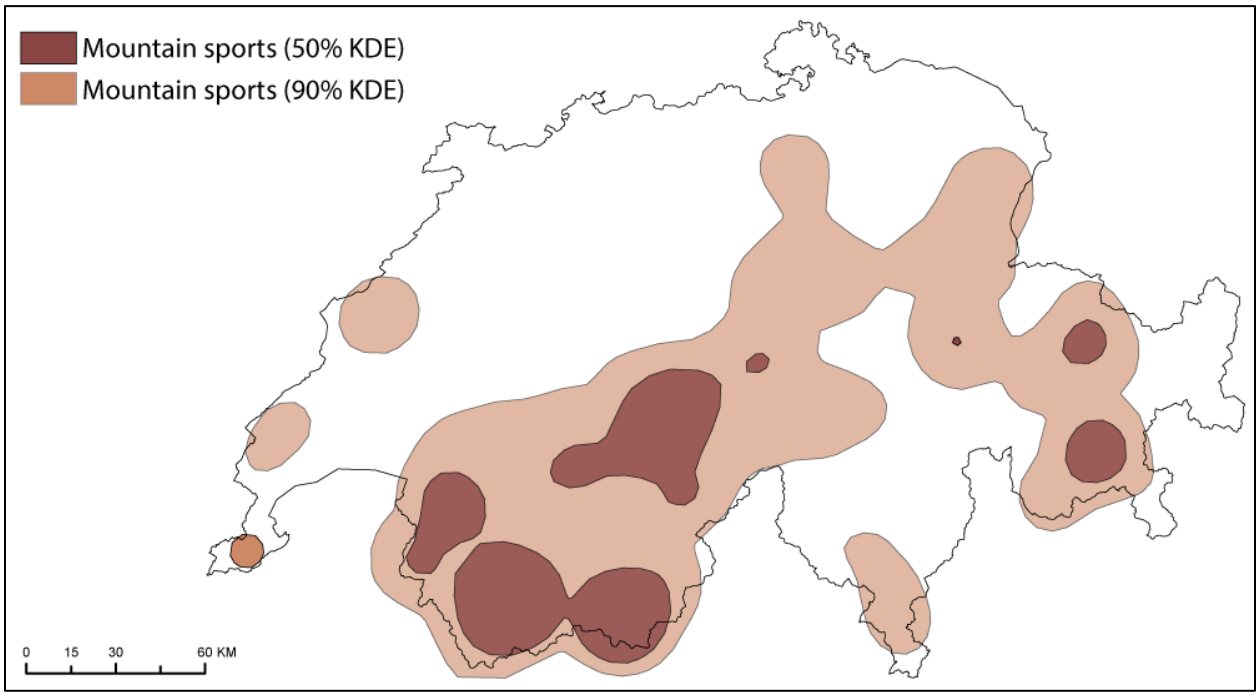


(b)

Figure 0.7 - Map showing the 50% and 90% kernel density estimate of winter (a) and summer (b) sports term occurrences in Flickr tags. Orange polygon represents area of summer sports and blue polygon represents area of winter sport activities. Bandwidth was selected as 20km.



(a)



(b)

Figure 0.8 : Map showing the 50% and 90% kernel density estimate of water (a) and mountain (b) sports term occurrences in Flickr tags. Brown polygons denote area of mountain related sport activities and blue polygons denote area of water related activities. Bandwidth was selected as 20km.

Table 0.2 - Table showing area of different KDE layers for different types of sports

Type of Sport / Layer	Summer	Winter	Water	Mountain
50% KDE area	5589.57 km^2	3583.84 km^2	2250.29 km^2	4517.58 km^2
90% KDE area	27176.34 km^2	23847.7 km^2	14568.26 km^2	21249.20 km^2

Table 0.3 - Table showing area overlaps of different KDE layers for different types of sports

Type of Sport / Area overlap	Summer and Winter	Water and Mountain
50% KDE areas	2773.07 km^2	0 km^2
90% KDE areas	19730.78 km^2	6751.41 km^2

For Figure 0.8 the activity terms relating to water and mountain sports were manually selected along with input by work done on the TRIPOD[22] project. For selecting mountain activity terms a digital elevation model (DEM) at a resolution of 250m was used. This model is called RIMINI, provided by Swisstopo. The raster contained elevations from 0 to 4807m for the whole of Switzerland. Terms occurring in elevations between 800m to 4807m were selected and the rest were deleted in order to delete outliers such as images of people "buying skiing equipment" or "start of skiing journey". This is because the lowest elevations of skiing resorts in Switzerland are around 800m. E.g. Atzmännig is a ski resort close to Zurich, in the area of the Alps situated at a height of 840m. On the Jura side, La Robella is another ski resort with its lowest point at 800m. Although, the results are very much in line with what is expected, the area of overlap between these layers is interesting and reported in Table 0.3 and Table 0.4. The difference between the areas of summer and winter activities is not very substantial. The winter areas are slightly higher as compared to the summer ones in the Alps area. This can be seen, for example, in the area around

[22] http://www.geo.uzh.ch/~aje/pdf/edwardes_purves_giscience2008_presentation_extendedabstract_colour.pdf

Davos and the area of Jura near Lake Geneva seen in Figure 0.7. In the second case, there is no overlap between the 50% KDE of water and mountain. Again, this is expected as it could possibly imply that the water activities that people are talking about are almost never or seldom located on mountain lakes or in the vicinity of mountains.

The maps in Figure 0.7 and Figure 0.8 are an example of grouping activities into clusters based on their time and feature. Another activity grouping was performed based on the semantics of activity terms. They were done using term co-occurrence and groups of 8 activities. Some of these activity group names were taken from the taxonomy of leisure activities from research performed in leisure studies by Tinsley and Eldredge (1995). The activity groups are summarised in Table 0.4; the main activity groups are travel, celebration, social, work, sport, food, competition and recreation.

Table 0.4 - Table showing different type of activity clusters along with some example activity terms and their corresponding colour-coding and icon in the map (Figure 0.9). Map icons CC BY SA Nicolas Mollet mapicons.nicolasmollet.com.

Colour	Activity cluster name	Examples activity terms
Green	Travel	Travel, tourism, tour, fly, holiday
Red	Celebration	Parade, festival, carnival, Christmas, etc.
Black	Social	Wed, religion, birthday, meet, etc.
Grey	Work	Work, construction, clean, etc.
Orange	Sport	Hike, ski, paraglide, dive, mountaineer, etc.
Purple	Food	Barbecue, drinks, beer, dinner, etc.
Yellow	Competition	Games, compete, race, etc.
Blue	Recreation	Birdwatch, paint, dance, etc.

The map in Figure 0.9 is an example of what is stated earlier in the motivation of this thesis. The motivation of this thesis is a change in approach of generalisation and to make it more activity and place driven instead of what it is now. It is possible that a tourist is more interested in the activity that can be performed in a place as opposed to the characteristics of the place itself. This map shows the highest activity performed in an area and its corresponding icon, colour coded according to the group it best fits into. Table 0.4 shows a few examples of some of the activities in the activity groups.

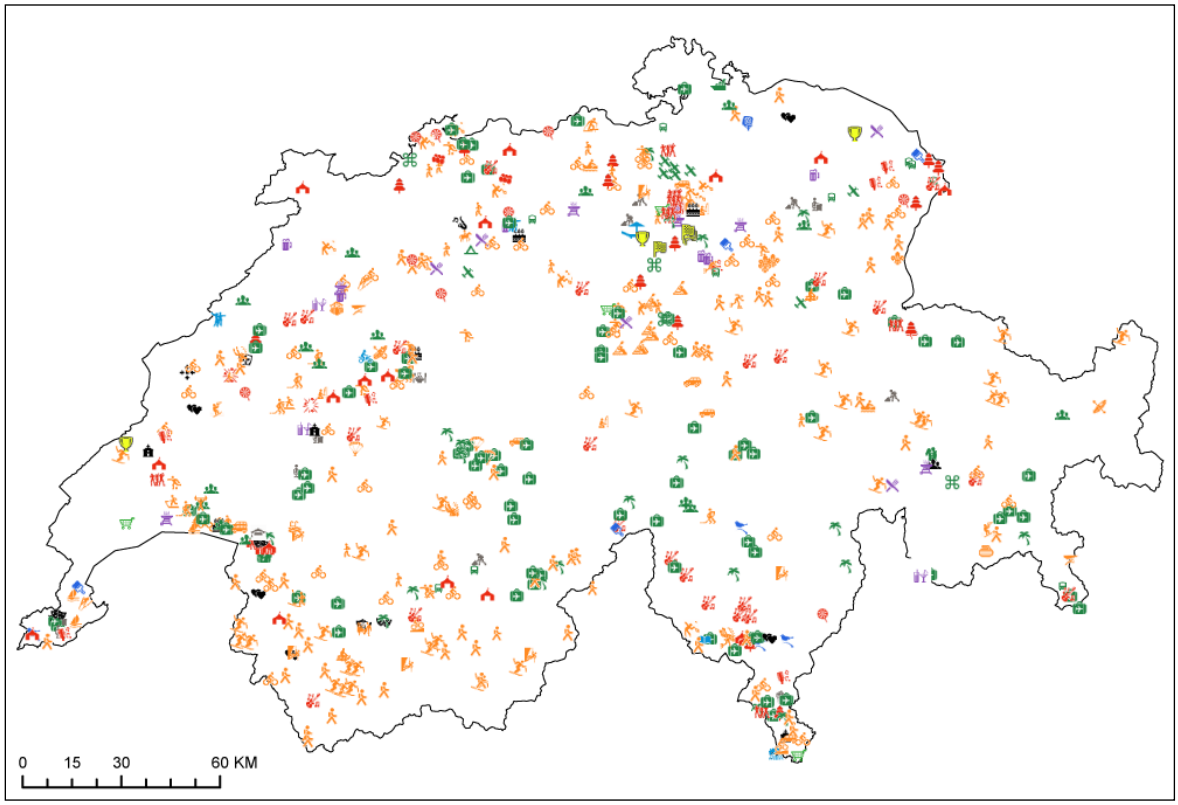


Figure 0.9 - Map of Switzerland showing various activity icons. Colour of the activity icon is based on the cluster it belongs to shown in Table 6.5 that shows the summary of the activity clusters and some examples of activity terms. Map icons CC BY SA Nicolas Mollet mapicons.nicolasmollet.com.

Discussion

1.25 RQ 1 – Web coverage across Switzerland

What is the web coverage across Switzerland for tourism related themes and how is it affected by different factors?

This research question pertains to the experiments and analyses performed in Chapter 4. In Chapter 4 the web coverage was measured through the number of web documents that exist for a given location. This metric is called the web count (Pasley et al., 2008). This experiment was performed for four languages, and various maps and analyses were presented.

1.25.1 Main contribution

The main contribution of this study are the maps and analyses that have been made to communicate that studies in geographic information retrieval have, thus far, tended to use quantitative web information about places for various decisions, and have assumed homogeneity of geographic coverage (Jones et al., 2008). However, the web coverage varies geographically and linguistically and is therefore not homogeneous, as has been shown by several studies. For instance, Graham et al. (2013) examine georeferenced tweets produced by Twitter users all over the world and plot a spatial tree map. This map clearly shows the inequality in the geography of content.

A method of normalisation is needed when dealing with quantitative analysis of web resources, since results could be biased by the amount of unequal data that exists for different places. Population is often used as a normalisation parameter, but this is not useful when experiments use only web content for their analyses, as results could be skewed due to the variation in content. Therefore, in such a case, web counts can be used as a variable for normalisation, in combination with population and touristic popularity. Large differences can also be attributed to location (e.g. proximity to city) and language. We show this in the graduated circle maps (Figure 0.4 and Figure 0.5) and by visualising the differences between the web coverage for four different languages (Figure 0.6), respectively. We are able to visually show how language causes bias to the extent that, for a given place, the amount of web content is sometimes very low for a particular language and very high for another (Figure 0.7). This means that while conducting research, the language in which it is conducted needs to be selected carefully, as the results could vary greatly based on

the selection. This is true especially for places that are multilingual. Li et al. (2013) use georeferenced Twitter and Flickr data to derive patterns rather than using only one of them, as they acknowledge that there is uneven distribution of the data generated in social media and the nature of such data has to be understood and used appropriately.

The discussion of the research related to web coverage (RQ 1) carries on in three parts with corresponding sub research questions. The first part discusses the general variation in the web coverage, along with a quick interpretation of the maps that indicate the variation. The second part discusses the variation due to language and a more detailed analysis of English, German, French and Italian. It also discusses the variation in three different datasets that were used to produce the web counts. Finally, the third part discusses two geographic factors and their correlations that may affect the geographic coverage.

1.25.2 Variation of Geographic web coverage

How does the geographic distribution of web coverage for tourism-related themes vary across Switzerland?

Findings and contributions

Using web counts is a relatively inexpensive method of measuring the background coverage of a particular collection and can be quickly carried out. Such an approach then allows the exploration of values which differ from the underlying distribution. The web counts are not only a measure of overall coverage, but also a fundamental first step before drawing conclusions based on the coverage of some specialised collection. The web counts are only a proxy of what really exists in terms of content regarding a particular theme. Nevertheless, inspection of the top twenty web pages in our case revealed that these pages most often contained web pages from the official website of the city, Wikipedia, Wikitravel, TripAdvisor, Qype, Yelp, MySwitzerland, Viator, Yahoo! Travel, and other similar websites which clearly do relate to tourism. The geographic distribution of these web counts seems to be most affected by language (Figure 0.6) and population (Figure 0.9). This can also be seen in Table 0.10 and Table 0.11, showing the top 10 web counts. These toponyms are often major cities and tourist resorts, and they can be seen clearly in the maps showing the hotspots (Figure 0.8 and Figure 0.9), which also suggests that there is a correlation between higher web counts and cities and their neighbouring places. Circular clusters of hotspots can also be seen for cities such as Zurich, Geneva, and Basel. This suggests that proximity of a place to a big city also seems to play a role in higher coverage. On the other hand the hotspots for Grindelwald, La

Chaux-de-Fonds and Davos are more linear. This pattern suggests that there are several distinct points of interest, rather than a cluster of points around a larger place (e.g. in Grindelwald area), or that alternatively the coverage depends on the terrain, e.g. for linear patterned hotspots in the valley surrounding Davos. The different coverage maps (Figure 0.4 and Figure 0.5) also show that the coverage is affected by the toponym datasets used. The valleys are better covered than the mountainous areas in both the SwissNames and GeoNames datasets. Big cities in general have larger counts and the area around the Alps in general has sparse coverage, but comparatively counts are higher in the SwissNames dataset.

Limitations

The raw web counts are only approximate values for measuring the coverage. This method does not account for artificially high occurrences of a toponym due to ambiguities or other reasons such as an event that caused the toponym to occur widely in the news, such as the World Economic Forum held in Davos. By virtue of the counting mechanism, the web counts convey aggregate coverage rather than individual trends. Furthermore, some toponyms were removed due to semantic ambiguities. The coverage of the search engine and that of the gazetteer also introduce a bias in our results, given that neither provides exhaustive coverage.

1.25.3 Language differences in the web coverage

Are there any differences in web coverage distribution for different languages and gazetteer datasets?

Findings and contributions

The web counts differ for different languages and this is seen very clearly in the graphs (Figure 0.1 and Figure 0.2) and coverage diagrams (Figure 0.4 and Figure 0.5). The observation that German is very well covered but Italian is not corresponds to the linguistic distribution of the Swiss population. On the other hand we see that the spatial autocorrelation is the least for English, translating into a wider coverage area and a tendency towards the coverage being more dispersed as compared to the other languages. French web counts, on the other hand, seem to have moderate coverage but are spatially highly correlated with the French-speaking region.

From the two bar charts (Figure 0.3), English and French show similar behaviour in both, the SwissNames and GeoNames gazetteer datasets, in terms of the cardinality of their web counts. However, after looking at the kernel density map (Figure 0.6), we can see a clear bias of French

web counts to the French speaking part. This is also the case for toponyms in the German and Italian speaking parts of Switzerland; they are better covered in the German and Italian languages, respectively.

The maps of χ values (Figure 0.7) show the comparison between the average web counts and the web counts in four languages, effectively highlighting the difference between expected and observed counts. Assuming that calculating the average web count is a way to reduce the bias caused by language, we are able to examine how much each language differs from average web counts. English, as mentioned earlier, seems to converge (lighter colours) more than the other languages, suggesting that coverage is more homogeneous than in other languages. This means that web pages on toponyms from Switzerland in the context of tourism are not concentrated on a particular language region in Switzerland. The spatial coverage in English is relatively homogeneous as compared to German, French and Italian. These languages show a bias towards the language region. In other words, there are more web pages about the region in the corresponding language, which make the spatial coverage relatively more heterogeneous for these three languages.

Limitations

Although the search engine queries were made in four different languages, there is a chance that even though these keywords occurred in webpages returned by the search engine, the contents may not be fully in the respective language. This in a way may lead to a false picture of the coverage. For instance, in English webpages it is common for people to use local toponym names for various reasons. A better approach, however, would be to use some language detection on the webpage by looking at the contents of the webpage instead of phrases.

1.25.4 Factors affecting the web coverage

How do typical factors such as population and touristic popularity of a place affect web coverage?

Findings and contributions

One might guess that the population of a place has a positive effect on the amount of web content for a given place. Highly populated places tend to have better transport infrastructure. For such places, more information that is important in the context of tourism is potentially available. On a similar issue, for instance, Graham et al. (2012; 2013) report through their visualisations (Figure

0.2) the digital divide in the geography of the Internet and the inequality in the geography of content (Twitter).

Considering Table 0.10, Zurich, Geneva, Bern and Basel are present in the top 10 web counts across all the languages, and they also happen to be the four most populated cities in Switzerland. However, when we computed the correlation between population and counts for places and cantons the results were not as expected. On further examination we noticed this behaviour could be due to a very large number of geo/non-geo ambiguities. These ambiguities cause the web counts to be artificially high for many tiny villages, which are otherwise not of interest to most tourists. Some examples of these cases are places named Wald (forest in German), Burg (castle in German), and Hard. For a more meaningful result, we computed the correlation (r) between places with the top 100 average web counts and their corresponding population. The result was 0.73, which hinted to a strong positive correlation. The places with top 100 average web counts were chosen after performing a manual disambiguation for all languages for these places.

To measure the popularity of a tourist destination is not straightforward. The web counts themselves do convey some information about the popularity of a place, but not explicitly. Hence, we selected the number of hotel nights rented per year per canton as a better indicator of touristic popularity and compared them to the web counts of the corresponding cantons through the method of correlation. We found that for French, English, and Italian the correlations of web counts to hotel nights per year are higher than web counts to population (Table 0.12).

Limitations

As mentioned in the paragraph above, we only have hints in the form of correlations from the factors that we hypothesised to affect the coverage the most; i.e. population and language. For a given place, the number of internet connections, its accessibility, spatial factors such as its terrain, daily commuter flows, public transport connections (especially in the case of Switzerland), and its vicinity to a big city or important touristic landmark could affect the coverage as well. Also, we have not directly examined any temporal factors such as the season or proximity in time to a big festival or event. It is possible that a certain toponym may have high counts because of the above reasons and has not been filtered out or dealt with.

1.26 RQ 2 – UGC and people's perceptions of places

This research question relates to the experiments and analyses performed in Chapter 5. In Chapter 5 we used Flickr data to examine people's perceptions of place and affordances from places in terms of the activities they perform in these places.

How can UGC be utilised to investigate people's perceptions of places? How can UGC be used to make affordances on a certain place?

1.26.1 Main contribution

The main contribution here is establishing the link between people's perception of place and the actions they are likely to perform in that place (Jordan et al., 1998). Actions, as discussed in the background chapter (Chapter 2), are measured by the activities people perform. Since this research focuses on people's perception of place, instead of using already defined place boundaries, such as administrative boundaries, we devised a method that is able to automatically extract locations from simple place-based terms in georeferenced Flickr tags. With this method, we are able to draw up areas and positions of toponyms across Switzerland, and since the method looks for terms of place in UGC, this can be seen as a reflection of people's perception. We gather activity terms for every location, thereby establishing a link between place and activity.

In the later part of this research (Chapter 6) we were able to generate maps of winter and summer activity clusters for Switzerland, as an example of 'activity landscapes'. These clustered areas could be treated as highly abstracted maps that can inform the generalisation process of the areas of perceived winter and summer activities respectively.

1.26.2 Locations and activity terms

With the help of UGC, can we extract locations of places and its activities? Is it possible to assign individual locations to these activities?

Findings and contribution

Research in Chapter 5 was based on the user's perspective of places and their boundaries that can be observed using UGC such as Flickr image tags. To associate activities with locations, the first step was to determine individual locations of places and commonly performed activities. Only then could a link be made between them. In line with the motive of this research we decided to devise

our own mechanism of location extraction from the UGC information instead of using administrative boundaries of places, which would have been detached from the perception of places ingrained in the UGC. To determine locations of places, we devised an automatic shrinking bounding box (ASBB) method whose workflow is described in Figure 0.1, with step-wise iterations shown in Figure 0.2. The location of a place could be approximated and/or represented using a range of geometries, such as a point, minimum bounding box, convex hull, or generalised polygon (Frontiera et al., 2008). If approximated well, the bounding box has been found to exhibit a balance of quality and simplicity to support GIR tasks (Janée and Frew 2004, Zhou et al. 2005). The approximation method used in this case was a combination of the distance between the bounding box centroid to the georeferenced point, which should be less than 10 km, and the size of the calculated bounding box, which must not be too big. Figure 0.3 shows the results of the ASBB method. These results give us an idea of how accurate the method is and based on these results, we filter out places. We also calculated the correlation of bounding box areas to the real areas of their corresponding locations, but did not find a high correlation to report. We suggest that this result is because people's perspectives of places are not proportional to what the actual administrative boundaries of a place are.

Limitations

One can argue that a bounding box may have not been the best and most accurate way of marking a location. In the past other ways to extract location from UGC have been proposed. Although the bounding box may not have been an ideal geometrical shape to plot for a location, it can be efficiently computed due to its simplicity (Frontiera et al., 2008). Furthermore, a seemingly more accurate representation such as a polygon has the disadvantage that it purports a degree of accuracy that is not really supported by the toponym denotations contained in the UGC. For various reasons, users frequently make errors in denoting a place, they denote a place when the picture was taken at a different location, or various forms of toponym ambiguities may occur, such as those discussed in Chapters 4 and 5. Nevertheless, an improved location extraction method should be able to deal with inaccuracies of location denotations, such as outliers, to which bounding boxes admittedly are very sensitive. A significant improvement would be to implement a rotated minimum bounding box or a convex hull, which both represent tighter approximations of a point set than an axis-parallel bounding box (Brinkhoff et al., 1993), or Alpha shapes (Edelsbrunner et al. 1983), which are even tighter but which also require more experimental parameterisation (and significantly more computational effort). Besides these polygonal representations, which do not reflect the distribution pattern of points within the polygon, representations that better reflect the ‘core’ of

the Flickr point locations could be examined. These include simple point representations such as a the mean or median centre, possibly accompanied by geometric distribution statistics such as the standard distance or standard deviation ellipse (Ebdon 1985), or density distributions such as kernel density estimate (KDE) surfaces (Hollenstein and Purves 2010), which can also be converted to polygonal representation by, for instance, computing the 90 % or 50 % contour of volume under the surface (Wartmann et al. 2010). Upon closer inspection, each representation method has its strengths and weaknesses. Thus, possibly a combination of different representations — each optimised for a particular application purpose — would be suitable. For the purposes of the present work, however, the bounding box was fully sufficient, as the aim of the experiment was to study activities in an area, rather than determining the area itself.

As mentioned above, the method of iterative shrinking of bounding boxes is affected by ambiguities. Therefore, ambiguities — both geo/geo and non-geo/geo — have been eliminated from the list of toponyms. Instead of complete elimination, basic methods such as those discussed by Rauch et al. (2003) and Buscaldi (2011) could be used to deal with ambiguities and then continue with the process, thereby yielding more locations.

1.26.3 Location similarity based on activities

Having extracted locations and their activities, is it possible to group these locations based on how similar the activities performed there are?

Findings and contribution

Initially, in the introduction of this thesis, the work was motivated by a scenario of a data enrichment process and by this means, generalisation algorithms could use information gathered on activities to generalise and thereby aggregate similar locations. The answer to the above research question works directly on a criterion to aggregate locations based on their similarity by computing the semantic similarity based on the activities and their frequency performed in particular places.

To take this further, we performed a cosine similarity analysis on our data. For all n bounding boxes, the activities along with their frequencies were listed, resulting in n vectors, and a similarity analysis was performed which resulted in an $n \times n$ matrix of values that ranged from 0 (no similarity) to 1 (full similarity, congruence). From the $n \times n$ matrix three vectors or locations — Zurich, Zinal and Lugano — were selected and listed in Table 0.4, along with the top 10 similar

locations. In Figure 0.11, Figure 0.12 and Figure 0.13 then, the three locations are further compared to their most similar 'partner' location (Table 0.1). For instance, Zurich and Adliswil share a similarity of 0.95 based on their activities. This can be attributed to the proximity of Adliswil to Zurich as seen in Table 0.1 and Figure 0.1. Adliswil is a typical suburban laid back residential area in close proximity to the city of Zurich and lies in the Canton of Zurich. Therefore activities such as 'party', 'parade', and 'music' are activities that have seeped in from the city of Zurich and are not really activities that are part of the town of Adliswil. The bounding box of Adliswil is also too large (as compared to its administrative boundary) and therefore gradually touches many of the areas in the city of Zurich (Figure 0.1). The activity of 'shopping' ranks quite high, as the bounding box of Adliswil contains Sihlcity, which is a well-known shopping mall on the outskirts of Zurich in the direction of Adliswil.

From the above result it is clear that big cities seem to have a seeping effect on smaller places that are in close proximity. The top 6 locations that are similar to Zurich are within 15 km of it. In terms of activities this also supports Tobler's First Law of Geography (Tobler 1970: 236). Referring back to the similarity map of Switzerland in Chapter 5 (Figure 0.11), an agglomeration of places similar to Zurich can be seen in the vicinity of Zurich, effectively visualising the behaviour discussed above.

Table 0.1 - Table showing the frequency of activity terms calculated from Flickr tags for a set of two most similar places. Similarity is grouped by colour.

Zurich	parade–520, travel–383, christmas–343, music–320, party–310, bike–288, shop–226, walk–199, concert–180, event–169, trip–152, holiday–138, sport–123, vacation–121, transport–108, bicycle–106, soccer–105, work–92, football–83, club–82
Adliswil	travel–242, christmas–214, party–163, parade–152, music–141, shop–138, hike–136, trip–105, holiday–94, vacation–86, walk–75, bike–75, concert–67, soccer–66, football–61, bicycle–51, festival–51, transport–47,sled–45,trek–42
Zinal	ski–61, hike–47, mountaineer–15, climb–10, travel–10, walk–10, rock–9, vacation–9, snowshoe–4, tourism–4, trip–4, bike–3, cycle–3, explore–3, fly–3, paraglide–3,
Arolla	ski–42, hike–41, vacation–13, rock–10, climb–6, mountaineering–4, travel–3
Lugano	travel–76, hike–46, bike–38, tour–35, holiday–31, shop–31, music–29, sailing–24, trip–23, explore–22, vacation–22, cycling–21, walking–21, christmas–20, rock–16, sport–16, tourism–15, mountainbike–14, newyearseve–12, festival –11
Morcote	travel–25, hike–24, sail–17, music–16, shop–13, bike–11, newyearseve–11, trip–11, christmas–8, walk–7, explore–6, tour–6, tourism–6, concert–4, festival–4, easter–3, ride–3, swim–3, vacation–3, wed–3

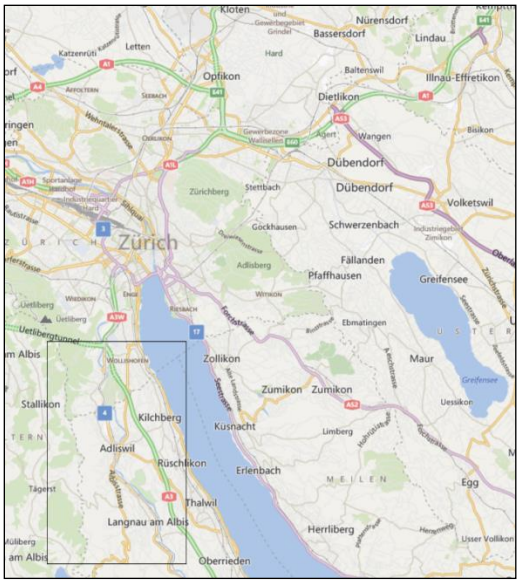


Figure 0.1 - Bounding box of Zurich (entire map) versus the bounding box of Adliswil

On the other hand, in the second group of similar locations in Figure 0.12, places such as Zinal and Flums are ski resorts well known for skiing, hiking, and other mountain activities. Both regions have more than one mountain peak and have well developed infrastructure to reach the tops of mountains. Once again, referring back to the similarity map of Switzerland in Figure 0.12, all places that are similar to Zinal are concentrated in the main chain of the Alps. Villages such as Arolla, Flums, Grimentz, Fiesch, and Engelberg have developed public infrastructure to access the mountains and well-developed hiking and skiing trails. Therefore, infrastructure and terrain also have a role to play in determining the type of activity performed, and consequently how semantically similar any two given locations are. The relationship between activity and infrastructure is discussed more in detail in the next section.

As mentioned above, the top 6 locations that are similar to Zurich are very close to the city of Zurich (Table 0.4). On the other hand Rheinfelden, Schaffhausen, Basel and Fribourg are cities, albeit smaller ones (with the exception of Basel) in other areas of Switzerland. It appears that the pattern of similarity of activities seems to be influenced by geographic distance and the infrastructure offered in a place. For places with administrative centres such as Zurich and Lugano, geographic distance appears to be more influential. For smaller places that are dominated by tourism-oriented infrastructure, of which Zinal is an example, the effect of geographic distance is

less pronounced. Hence, the places that are similar to Zinal are small Swiss towns that house ski and hiking resorts, spread over all of the Alps, as opposed to the immediate vicinity.

Limitations

As inferred from the literature, the cosine similarity can inform the semantic ranking in the case of toponyms, given it provides a measure of 'relatedness' as often used in GIR (Hecht et al., 2012). Even though the results seem very much in line with the geography of places, the similarity is, after all, computed based on terms in natural languages. The term may very well mean something else, and an improvement would be to look at the context of the term in the whole Flickr tag instead of the independent term itself. For instance, 'skiing' and 'snowboarding' are activities in Zurich that rank 25th and 28th, respectively, in the frequency list of activities. Realistically speaking, these activities are not often performed in the bounding box of Zurich. On examination most of these images were photographs taken by people commencing or ending a skiing or snowboarding trip, such as pictures taken in the train while travelling for such an activity. There were also pictures taken of people buying skiing and snowboarding equipment in sports shops.

1.27 RQ 3–Activities in space and time

Once extracted, how do these activities behave spatially?

This research question relates to analyses performed in Chapter 6.

1.27.1 Main contribution

Activity infrastructure was examined for meaningful correlations, particularly for the activity of 'hiking' and other similar terms. Activity terms were then analysed by applying the first order co-occurrence between each other. Co-occurrence brings out the semantic similarity in terms. We used it to examine the interaction between activities. Finally, a temporal examination of activities was made, enabling us to identify temporal patterns, thereby automatically detecting temporal clusters of activities based on seasons.

The first contribution mentioned above will be discussed in more detail in Section 1.27.2, while the discussion of the remaining three contributions follows in Section 1.27.3.

The map of Figure 6.9 reported in Chapter 6 is an example of the outcome of the motivation discussed earlier on. It is an example of the potential of a method of generalisation that focuses on what activity can be performed in a place and not the importance or size of the place *per se*.

1.27.2 Activities w.r.t. space and topographic data

How do these-tourism related activities relate to space and how do they relate to topographic data?

Findings and Contributions

In the previous section (1.26.3), one of the conclusions drawn was that infrastructure could have a positive effect on activities performed in a particular place. Section 1.21 analysed this observation for the activity of hiking. Hiking was selected as it is a very popular activity in Switzerland and it is common among both locals and tourists. We found it to be the second most popular activity (term) in Switzerland (Table 0.3). This activity was examined against hiking infrastructure that is, hiking trails for Switzerland. Switzerland was divided into a grid of 10 km by 10 km and for every square the number of hiking points was compared to the length of the total trail. The correlation was calculated between them using Pearson's r, first for all grid cells and the points contained in them, and then only for grid cells that contained at least one hiking related Flickr point. Understandably, the correlation went up in the second case.

To further illustrate this argument, we examined the bounding boxes of Zinal and Flums. Flums is a village that hosts a large ski resort known as Flumserberg, therefore Flumserberg was examined instead. It is located in the Canton of St. Gallen. Zinal is a resort in the area of Val d'Anniviers in the French speaking part of the Canton of Valais. These two boxes have been rated very similarly based on the activities that people perform (Table 0.4). Table 0.2 shows the comparison between these two places in terms of other activity infrastructure present, from which it is clear that these places are indeed very similar in terms of the infrastructure offered, as both are alpine resorts. This information was obtained from the official tourism website of the two resorts. Other than the basic infrastructure for skiing, hiking, etc., these resorts also have supporting infrastructure such as rental shops for skiing and other sports equipment, parks for children, picnic areas, climbing areas, hotels, cafes, and mountain huts. Looking back at Chapter 4 the relative web counts for these two areas are also quite similar. In the case of Flumserberg it is 28,600 and 19,700 in the case of Zinal. This means that the areas of Zinal and Flums are expected to be quite similar, and indeed the similarity values are high for both activities and infrastructure.

Table 0.2 - Table showing infrastructure of Zinal[23] and Flumserberg[24] ski resorts[25]

Infrastructure	Flumserberg	Zinal
Hiking trails	~150km	~300km
Bike trails	~125km	~150km
Skiing paths	~65km	~70km
Cross country ski trails	~18km	~19km
Sledging trail	~3km	n/a
Toboggan run	2km	200m
Highest altitude	2222m	2895 m
Lowest altitude	1000m	1670 m
Number of ski lifts	18	19
Number of ski schools	1	1

Limitations

The initial research question deals with the relation of activities to space. Infrastructure is only one possible explanatory spatial variable, and our initial guess was that it relates well with activities. However, there are other variables too, such as terrain, ease of accessibility, public transport infrastructure, and proximity to a large city. Although these variables have not been examined, we expect to find meaningful correlations with some of them.

The correlation calculated for the grid cells is only approximate, and in reality might be higher.VECTOR25 provides exhaustive information on hiking trails in Switzerland. We compare these trails to hiking points in Switzerland on Flickr. Hiking points are extracted from Flickr by filtering photos that contain tags with terms ‘hike’, ‘walk’, ‘trek’ and “mountaineer”. These points are not exhaustive as Flickr is not an exhaustive dataset of what people do and the georeferenced set of images from Flickr is an even smaller percentage. Therefore, while calculating the correlation, we are comparing two unequally covered datasets, one of which is exhaustive while

[23] http://www.valdanniviers.ch/tourism/zinal-ayer.html

[24] http://www.flumserberg.ch

[25] MySwitzerland.com (Official tourism webpage for Switzerland)

the other is not. A better solution to this problem would be to first automatically georeference the complete set of Flickr images for Switzerland (Wider et al., 2013) and then repeat the analysis.

1.27.3 Activity interaction and behaviour

How can interaction between activities be measured, what is the nature of their interaction and how do they behave spatially and temporally?

Findings and Contributions

In the case of activities, their interaction was measured through co-occurrence of terms, which reports the semantic proximity of terms in a corpus, and thus the semantic similarity (Section 1.22). Co-occurrence is a well-established similarity measure in linguistic studies and has also been applied in GIR (Bordag, 2008). Computing the co-occurrence of two activity terms may tell us about the potential interaction of these terms in a corpus. The corpus used here consisted of tags in Flickr photographs. The window of text used in the co-occurrence computation is an important parameter in linguistics, and clusters are known to be impacted based on the window of text being an entire document, a sentence, a set of sentences, or fixed-length window of text (Momtazi et al., 2010). In our analysis two windows have been used: a window per individual user and a temporal window per user per day. In both cases the results and the co-occurrence ranks are quite similar. For instance, the top ten activity pairs for both cases have eight pairs in common (Table 0.1). We conclude from this that — at least in our case — the choice of temporal window is not important: The nature of an activity causes it to be more similar to another activity, rather than the window of time over which these activities are being performed.

The next part of the discussion concentrates on an analysis that was performed in order to examine the pattern of the activities owing to temporal and spatial aspects (Section 1.23). All records that are downloaded from Flickr have multiple attributes such as ID, name, date taken, date posted, and tags. We examined the date taken and date posted attributes for temporal patterns, reported in Figure 0.4, Figure 0.5 and Figure 0.6. These results allowed making interesting observations regarding the temporal pattern of activities. Similar to the work of Edwardes and Purves (2008), we were able to make out summer and winter activity term clusters.

Using these term clusters — i.e. summer, winter, water, and mountain activities — we generated 50 % and 90 % kernel density contours of their spatial distribution (Section 1.24). 50 % and 90 % were chosen as these areas have been often used as an estimate of the home range and the core

area, respectively, of animals in behavioural studies (Wartmann et. al, 2014). The result is a highly abstracted map of which activity-high areas are known.

The last part of this research ends with fulfilling a part of the motivation: to create an activity-based, generalised map. Figure 0.7 and Figure 0.8 show examples of how, by extracting activities from UGC, generalisation in the sense of abstraction can be achieved, creating maps of summer vs. winter sports (Figure 0.7) and mountain vs. water sports (Figure 0.8), respectively. Figure 0.9 then provides an example that demonstrates the potential of UGC as a driver of map generalisation, by semantic enrichment of spatial data. Although no geometric generalisation has been applied to the points depicted, this map shows how activities extracted from UGC form enriched place descriptions, thus providing information that can be used to influence adaptive generalisation, cases of which are discussed in Bereuter et al. (2012). Adaptive generalisation in this context can create personalised maps of activities for users depending on their profile and interests. There has been research on egocentric maps and various methods of geovisualisation (Meng, 2004; 2011) and relevance ranking for mobile users (Crease and Reichenbacher, 2011; De Sabbata and Reichenbacher, 2012; Reichenbacher, 2003;2009) concerning everyday activities. However, so far activities have not been used to directly inform visualisation and ranking, and in turn the creation of generalised, personalised maps.

Limitations

Activity interaction was studied through activity co-occurrence. Co-occurrence in linguistic studies also uses more advanced methods such as latent semantic analysis (Maletic and Marcus, 2000) and high order co-occurrences (Lemaire and Denhière, 2006) to dig deeper into term similarity, thus identifying patterns and evolving relationships between terms.

In our temporal analysis, the ‘date taken’ and ‘date posted’ parameters can only approximately say something about when a picture was taken. That is, there is a possibility that a user's camera may not have had the right time stamp on it due to various reasons. Similarly, it is possible that a user posted his/her pictures much later than when they were actually taken.

Although a couple of activity clusters were made through basic analysis of temporal events, it might also be interesting to apply traditional clustering to extract more patterns. After extracting activity terms from Flickr it would be possible to create a taxonomy of sports activities based on a number of criteria, such as geographic feature related (e.g. mountain activities), season related, etc. As a first step, applying an agglomerative hierarchical clustering algorithm to these activity

terms will yield a basic set of hierarchies, which with some improvement can feed the basics of taxonomy of sports related activities. Alternatively, partitioning clustering approaches such as k-means could be applied to the group of activities, to yield groups of activities that are similar w.r.t. particular features.

So far in the literature, grouping of text mined activities based on spatio-temporal criteria such as relation to geographic feature and/or variations over seasons, has not been explored in detail. This opens interesting questions such as computing activity similarity and exploring and applying various criteria to compute this.

Conclusion

This thesis has presented research carried out on methods and analysis for information retrieval in the geographic context. The main motivation of this thesis was to enhance the process of adaptive generalisation by providing supporting information, not available in standard MRDBs, through the use of user-generated content. Standard MRDBs often contain static map related data, therefore, the motivation was to build a supporting auxiliary database termed the FactsDB that is connected to the MRDB through a gazetteer. The research work does not propose the structure of the database, instead it proposes what kind of information can be extracted and stored in the FactsDB. Work reported in Chapters 4 to 6 talks about what kind of information could be extracted for that purpose. So far, only little work existed in the area of using UGC to improve map generalisation (e.g. Gaffuri, 2011; Huang and Gartner, 2012).

The next few sections summarise the main achievements, insights and open issues of this work and discuss directions for future research, along with an extension project related to this thesis.

1.28 Achievements and insights

The main achievements of this thesis are listed below and the section will be discussed in light of these points

1. Identifying that the web coverage varies geographically and linguistically.
2. Automatic shrinking bounding box approach for determining locations of populated places
3. Linking of these locations and activities performed in them, using UGC
4. Location similarity based on activities
5. Activity landscapes and activity maps of Switzerland

Geographic and linguistic web coverage

The first achievement of this study is a repeatable method using web counts as a proxy for the web coverage. This was performed to examine the coverage and variations caused due to language

(Chapter 4). Web counting has been used in the past (Kilgarriff and Grefenstette, 2003; Pasley et al., 2008) and in this case was used as a proxy for the coverage (Venkateswaran et al., 2014).

Each toponym from a gazetteer was queried using the Yahoo Boss API and the returned web count i.e., the number of web pages, was noted as the web count. In order to eliminate the bias caused by the gazetteer itself, places from two different gazetteers and tourist points of interest dataset were utilised. The study also discusses the effects of alternative toponym spellings (Table 0.10 and Table 0.11) and shows that the counts change at an individual level, but the aggregate results are not very different. Therefore, researchers should keep in mind the spatial extent of the study area when using this method.

When dealing with data in GIR, a method of normalisation is always needed while performing any kind of quantitative analysis of web resources, as these results could be biased by the amounts of unequal data that exist for different places. Web counts for toponyms is the Population is often used as a normalisation parameter, for GIR related studies, but this is not useful when experiments uses web content for its analysis. Therefore in such cases, the web counts are an ideal parameter that can be used for normalisation of these results. Not only for purposes of normalisation, but web counts are also an important metric in the context of any language examination. In the past web counts were used as a metric to build web based models for several natural language processing tasks in different languages and to approximate bigram counts (Keller and Lapata (2003); Lapata and Keller (2005)).

Automatic location determination allows linking activities to named locations

Through the automatic shrinking box approach (ASBB), using UGC, boundaries around places are automatically generated for approximately 2500 toponyms form Switzerland. To associate activities with named locations, the first step was to determine individual locations of places and commonly performed activities, only then a link could be made between them. Since this research focused on people's perception of place (Cresswell, 2009), using already defined place boundaries (e.g. administrative boundaries) would have defeated the purpose of this research. Flickr points containing image metadata were extracted from 2006 to 2011 for Switzerland and the tags were examined for toponyms of populated places from the SwissNames database. For every toponym a set of Flickr images was made and a shrinking box with the best fit was generated automatically. For each bounding box of populated places the frequency of activity terms in Flickr tags was

calculated. The link between place and activity could then be established as these same pictures on Flickr contain both the place term and the activity term.

Location similarity based on activities

One of the motivations of this research was to create a supporting database known as the FactsDB, through which the spatial data of the MRDB could be enriched with additional information, which at a later stage could inform the generalisation process, providing additional semantics about particular places. One particular piece of information that is a candidate for being part of the FactsDB is the semantic similarity between different places, which would allow aggregating them into classes of places linked to similar activities. In this thesis, semantic similarity between places was computed based on the activity frequencies using the cosine similarity method.

It can be hypothesised that infrastructure also plays an important role in the type of activity performed. Simply put, certain types of infrastructure afford certain types of activity. Therefore, hiking trails were examined against hiking points to find whether a positive correlation existed between the two that was higher in the mountainous regions. As the results presented in the discussion showed, for the case of skiing resorts infrastructure similarity is also high for places that have a high semantic similarity.

Both the above achievements directly link back to the motivation of a more adaptive method of generalisation. Information on the bounding box geometry could be stored in the FactsDB, denoting the user's perspective of a place. Along with this similar generalisation approaches can be applied to locations exhibiting similarity due to parameters that can be decided on-the-fly. This means that locations could be similar because of their population, size, infrastructure, and as discussed above, activities.

Activity landscapes and activity map of Switzerland

Activity landscapes in this research are equivalent to core activity areas for Switzerland. For example, we were able to generate maps of winter, summer, water and mountain activity clusters for Switzerland, as an example of activity landscapes. The seasonal clusters were identified through a temporal grouping analysis and the water and mountain sports clusters, respectively, were identified from previous research on terms that produced all terms associated with water and mountain sports (Edwardes and Purves, 2008). These clustered areas were generated using 50 % and 90 % kernel density contours. The resulting maps could be seen as highly abstracted maps,

representing a generalised view of the areas of winter and summer activities, as perceived through the eyes of the contributors of UGC (i.e. Flickr) content.

Finally, the linking of activities to previously extracted locations also allowed creating an activity map of Switzerland with activity icons. Different colours of the icons were used to denote different activity groups. These groups were identified using a combination of activity co-occurrence and the activity taxonomy of Tinsley and Eldredge (1995).

1.29 Open issues

While this work has led to a number of achievements and insights, a number of issues have not been addressed and are discussed below.

Variation of geographic web coverage:

- The methods we used to examine the web coverage do not deal with the temporal variation of the counts. Therefore it might be possible that certain toponyms may have high or low counts because of the season or time proximity to a big event. For instance, the World Economic Forum took place in Davos from the 27th to 31st of January 2010 and the counts were gathered in February 2010. While the event was over, the media were still talking about this event and so were different bloggers around the world. Although Davos is a famous skiing destination and thereby a tourist spot, we think its counts could be artificially high due to this reason.
- The web counts chapter examines population density and tourism popularity as factors that could contribute to explaining the web coverage of a place. However, there are likely more than two factors that are of importance. For a given place, the number of internet connections, its accessibility, spatial factors such as its terrain, daily commuter flows, public transport connections (especially in the case of Switzerland) and its vicinity to a big city or important touristic landmark could affect the coverage as well.

UGC and people's perceptions of places:

- In order to ascertain boundaries of places using UGC (Fickr), we devised a method known as Automatic Shrinking Bounding Box and thereby created a bounding box to approximate an area around a place. The issue with this approach is that the output of this method is a

rectangular region which is not an accurate and differentiated way to denote place boundaries, as a minimum bounding rectangle is susceptible to outliers (thus systematically overestimating the area) and does not represent the density variation of the original UGC points. Alternative approaches such as the convex hull or a rotated minimum bounding rectangle may have been a better geometric shape to draw a boundary. Similarly, kernel density methods would allow to represent varying UGC point densities associated with a place.

Characteristics of the data used:

- The experiments performed make use of UGC, especially Flickr. Flickr is a representative of a very small community of users compared to the world's population. Therefore, the perceptions of places inferred from the corpus, reflect the perspectives of Flickr contributors, which has been extrapolated to people's perception. Flickr also suffers from participation inequality (Neilsen, 2006). Purves et al., (2011) calculate this inequality, stating that "73 % of the images are contributed by 10 % of the users". We tried different measures to eliminate this bias in Chapter 5, but some of the bias still remained.

1.30 Outlook and Future work

Extension of the web coverage research

As discussed above many parameters have been listed that could affect the web coverage. Using these parameters as explanatory variables it is possible to build a multiple regression model to predict the web counts and therefore the coverage. This is an alternative method to gathering actual web counts using a search engine. Search engine web counts could then be used to calibrate the regression model.

Using activity terms for disambiguation

It is possible to reverse engineer ambiguity based on the combination of toponyms and activity terms, especially in the case of geo-geo ambiguities. Activities are closely related to the features of a place, and from the discussion chapter is it clear that infrastructure in a place affords certain types of activities. Therefore, using Flickr or other UGC that is similar in nature, the question is whether it is possible to examine terms that contain a combination of toponyms and activities, which give

us hints about the type of activities that could be performed, thereby offering a disambiguation mechanism for geo-geo ambiguities.

Activity Ontology

Research performed in the context of activities, gives way for a more organised methodology to classify activities and therefore a resulting taxonomy or ontology of activities. A simple ontology of tourist activities or sports activities can be carried out, in line with work done by Tinsley and Eldredge (1995). Kuhn (2001) works on a similar concept, except that the activities extracted are in the context of driving. As Alazzawi et al. (2012) suggest it should also have a temporal aspect as the type of activity performed in a place may vary with time. An effective taxonomy will also make it easier to conduct research across languages.

Project PlaceGen

With some inputs from this work, a project named PlaceGen (Place-based Map Generalization) was proposed by the advisors of the author and work on this project began in July 2014 with two new PhD students. The objective of this project is to develop methods to make use of place descriptions extracted from UGC to achieve a new form of map – place-based maps. Such maps can be accomplished by linking these descriptions to drive adaptive generalisation, thereby achieving place-based generalisation.

Bibliography

Aberley, D. and Sieber, R. (2002): Public Participation GIS (PPGIS) Guiding Principles First International PPGIS Conference. Held by URISA at Rutgers University, New Brunswick, New Jersey, 20—22 July.

Alazzawi, A.N., Abdelmoty, A.I. and Jones, C.B. (2012). What can I do there? Towards the automatic discovery of place-related services and activities. *International Journal of Geographical Information Science*, 26(2):345-364.

Ames, M. and Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 971–980, San Jose, USA.

Amitay, E., Har'El, N., Sivan, R. and Soffer, A. (2004). Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280, Sheffield, UK.

Ballatore, A., Bertolotto, M. and Wilson, D.C. (2012). Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowledge and Information Systems*, 37(1):61–81.

Bayardo, R.J., Ma, Y. and Srikant, R., (2007). Scaling up all pairs similarity search. In: *Proceedings of the 16th international conference on World Wide Web*, pages 131–140, Banff, Canada

Bereuter, P. and Weibel, R.(2012). Assessing Real-Time Generalisation of Point Data. In *16th Workshop on Progress in Generalisation and Multiple Representation map generalization*, Dresden, Germany.

Berners-Lee, T. and Cailliau, R. (1990). WorldWideWeb: Proposal for a HyperText project. European Particle Physics Laboratory (CERN).

Bishr, M. and Janowicz, K. (2010). Can we trust information?-the case of volunteered geographic information. In Devarahu et al. (eds) *Towards Digital Earth Search Discover and Share Geospatial Data Workshop at Future Internet Symposium*, Workshop at Future Internet Symposium, Berlin, Germany.

Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E. and Yergeau, F. (1997). Extensible markup language (XML). World Wide Web Journal, 2(4):27–66.

Buettner, M., Prasad, R., Philipose, M. and Wetherall, D. (2009). Recognizing daily activities with RFID-based sensors. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 51–60, Orlando, USA.

Bundesamt für Landestopografie, swisstopo. (Federal Office of Topography, swisstopo). http://www.swisstopo.admin.ch.

Bundesamt für Statistik Neuchâtel. (Federal Statistical Office, Neuchâtel). http://www.bfs.admin.ch/bfs/portal/en/index.html.

Buscaldi, D. (2011). Approaches to disambiguating toponyms. *SIGSPATIAL Special*, 3(2):16–19.

Buscaldi, D. and Rosso, P. (2008). A conceptual density based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, 22(3):301–313.

Butler, D. (2006). Mashups mix data into global service. *Nature*, 439(7072):6–7.

Canter D. (1997). The facets of place. In Moore G.T. and Marans R.W. (eds.), *Advances in Environment, Behavior, and Design, Vol. 4: Toward the Integration of Theory, Methods, Research, and Utilization.*, 109–147, Springer.

Carver, S., Evans, A., Kingston, R. and Turton, I. (2000). Accessing Geographical Information Systems over the World Wide Web: Improving public participation in environmental decision-making. *Journal of Information Polity*, 6(3):157–170.

Couclelis, H. and Getis, A. (2000) Conceptualizing and measuring accessibility within hysical and virtual spaces. In D. G. Janelle and D. C. Hodge (eds.) *Information, Place and Cyberspace: Issues in Accessibility*, 15-20, Springer.

Crandall, D.J., Backstrom, L., Huttenlocher, D. and Kleinberg, J. (2009). Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, pages 761–770, Madrid, Spain.

Cresswell, T. (2009). Place. Thrift., N. and Kitchen, R. editors, *International Encyclopedia of Human Geography*. Elsevier, 169–177.

Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V. (2002). {GATE}: A Framework and Graphical Development Environment for Robust {NLP} Tools and Applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, USA.

Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N. and Weerawarana, S. (2002). Unraveling the Web services web: an introduction to SOAP, WSDL, and UDDI. *Internet Computing, IEEE*, 6(2):86–93.

Curry M.R. (1996). *The Work in the World - Geographical Practice and the Written Word*. University of Minnesota Press, Minneapolis.

Daugherty, T., Eastin, M.S. and Bright, L. (2008). Exploring consumer motivations for creating user-generated content. *Journal of Interactive Advertising*, 8(2):1–24.

Dearman, D. and Truong, K.N. (2010). Identifying the activities supported by locations with community-authored content. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 23–32, Copenhagen, Denmark.

De Longueville, B. D., Luraschi, G., Smits, P., Peedell, S. and Groeve, T. D. (2010). Citizens as sensors for natural hazards: A VGI integration workflow. *Geomatica*, 64:41-59.

Derungs, C., Purves, R.S. and Waldvogel, B. (2011). Toponym disambiguation of landscape features using geomorphometric characteristics. In *Proceedings of the 11th International Conference on GeoComputation*, pages 106–110, London, UK.

Di, L., Zhao, P., Yang, W., Yu, G. and Yu, P. (2005). *Intelligent geospatial web services. In Geoscience and Remote Sensing Symposium, 2005. IGARSS '05. Proceedings*, Volume 2, 1229–1232. IEEE.

Edwardes, A.J. (2007). Re-placing Location: Geographic Perspectives in Location Based Services. Ph. D. Dissertation. University of Zurich

Edwardes, A.J. (2009). Geographical perspectives on location for location based services. *In LOCWEB '09: Proceedings of the 2nd International Workshop on Location and the Web*, pages 1–4, Boston, USA.

Edwardes, A.J. and Purves, R.S. (2007). A theoretical grounding for semantic descriptions of place. In *Proceedings of the 7th international conference on Web and wireless geographical information systems*, pages 106–120, Cardiff, UK.

Edwardes, A.J. and Purves, R.S. (2007). Eliciting concepts of place for text-based image retrieval. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 15–18, Lisbon, Portugal.

Egenhofer, M.J. (2002). Toward the semantic geospatial web. In *Proceedings of the 10th ACM international symposium on Advances in geographic information systems*, pages 1–4, McLean, USA.

Egenhofer, M. J., and Mark, D. M., 1995. Naive Geography. In Frank, A. U. and Kuhn, W., editors, *Spatial Information Theory: A Theoretical Basis for GIS, Lecture Notes in Computer Sciences,* 988:1-15.

Elmes, G., Dougherty, M., Challig, H., Karigomba, W., McCusker, B. and Weiner, D. (2005). Local knowledge doesn't grow on trees: Community-integrated geographic information systems and rural community self-definition. *Developments in Spatial Data Handling*, 29–39.

Fisher, P. and Unwin, D. (2005). Editors. *Re-presenting GIS*, John Wiley & Sons.

Flanagin, A.J. and Metzger, M.J. (2008). The credibility of volunteered geographic information. *GeoJournal*, 72(3):137–148.

GeoNames. www.geonames.org.

Gibson, J. J., (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton-Mifflin)

Girardin, F., Fiore, F.D. and Blat, J. (2007). Understanding of tourist dynamics from explicitly disclosed location information. In *Proceedings of the 4th International Symposium on LBS and Telecartography* (Hong-Kong, China, 2007).

Girardin, F., Fiore, F.D., Ratti, C. and Blat, J. (2008). Leveraging explicitly disclosed location information to understand tourist dynamics: a case study. *Journal of Location Based Services* 2(1): 41-56.

Girres, J-F. And Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset, *Transactions in GIS*, 14(4):435-459

Golledge, R., and R. Stimson. (1997). Spatial behavior: A geographic perspective. New York, New York: The Guilford Press

Goodchild, M. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69 (4):211–221.

Graham, M., Hale, S. and Stephens, M. (2012). Digital divide: the geography of Internet access. *Environment and Planning A*, 44 (5):1009–1010.

Graham, M., Stephens, M. and Hale, S. (2013). Mapping the geoweb: a geography of Twitter. *Environment and Planning A*, 45 (1):100–102.

Hahmann, S. and Burghardt, D. (2011). Maple – a Web Map Service for Verbal Visualisation using Tag Clouds Generated from Map Feature Frequencies. In A. Ruas (Ed.), *Advances in Cartography and GIScience. Volume 1*, 3–12.

Haklay, M., Singleton, A. and Parker, C. (2008). Web Mapping 2.0: The Neogeography of the GeoWeb. *Geography Compass*, 2 (6):2011–2039.

Hall, M.M. and Jones, C.B. (2008). A field based representation for vague areas defined by spatial prepositions. In *Methodologies and Resources for Processing Spatial Language, Workshop at LREC'2008*, pages 36-41, Marrakech, Morocco.

Haklay, M. and Weber, P. (2008). OpenStreetMap: User-Generated Street Maps. *Pervasive Computing*, IEEE, 7 (4):12–18.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37 (4):682 – 703.

Hampe, M. and Sester, M. (2002). Real-time integration and generalization of spatial data for mobile applications. Geowissenschaftliche Mitteilungen:167–175.

Hecht, B., Carton, S. H., Quaderi, M., Schöning, J., Raubal, M., Gergle, D. and Downey, D. (2012). Explanatory Semantic Relatedness and Explicit Spatialization for Exploratory Search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 415–424, Portland, USA

Hill, L.L. (2000). Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, pages 280–290, Lisbon, Portugal.

Hill, L.L. (2006). Georeferencing: The Geographic Associations of Information. MIT Press, Cambridge, MA

Hollenstein, L. and Purves, R.S. (2010) Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science* 1(1):21-48

Hollenstein, L. (2008). Capturing Vernacular Geography from Georeferenced Tags. University of Zurich

Honda, K., Hung, N.D. and Shimamura, H. (2006). Linking OGC Web Services to Google Earth. In *SICE-ICASE, 2006. International Joint Conference*, pages 4836–4839, Busan, Korea.

Jain, S., Seufert, S. and Bedathur, S. (2010). Antourage: mining distance-constrained trips from Flickr. In *Proceedings of the 19th international conference on World wide web*, pages 1121-1122, Raleigh, USA.

Jankowski, P., Andrienko, N.V., Andrienko, G.L. and Kisilevich, S. (2010). Discovering Landmark Preferences and Movement Patterns from Photo Postings. *Transactions in GIS* 14(6): 833-852.

Jiang, B. and Yao, X. (2006) Location-based services and GIS in perspective. *Computers, Environment and Urban Systems,* 30(6):712-725.

Jones, C.B. and Purves, R.S. (2008). Geographical information retrieval. *International Journal of Geographical Information Science*, 22 (3):219–228.

Jones, C.B., Purves, R.S., Clough, P.D. and Joho, H. (2008). Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, 22 (10):1045–1065.

Jordan, T., Raubal, M., Gartrell, B. and Egenhofer, M.J. (1998). An Affordance-Based Model of Place in GIS. In *8th Int. Symposium on Spatial Data Handling (SDH'98)*, pages 98–109, Vancouver, BC.

Joshi, D. and Luo, J. (2008). Inferring generic activities and events from image content and bags of geo-tags. *In Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 37-46, Niagra Falls, Canada.

Kaptelinin, V. (2012). Activity Theory. In Soegaard, M., and Dam, R. F. (2012). *In The Encyclopedia of Human-Computer Interaction*, MIT, Massachusetts, USA.

Keller, F. and Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29 (3):459–484.

Keßler, C., Janowicz, K. and Bishr, M. (2009). An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 91-100, Seattle, USA.

Khodaei, A., Shahabi, C. and Li, C. (2012). SKIF-P: a point-based indexing and ranking of web documents for spatial-keyword search. *Geoinformatica*, 16 (3):563–596.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29 (3):333–347.

Kisilevich, S., Krstajic, M., Keim, D., Andrienko, N. and Andrienko, G. (2010). Event-Based Analysis of People's Activities and Behavior Using Flickr and Panoramio Geotagged Photo Collections. In: *Proceedings of the 14th International Conference Information Visualisation*, pages 289-296, Los Alamitos, USA.

Kreveld, M. Van, Reinbacher, I., Arampatzis, A. and Zwol, R. Van. (2005). Multi-Dimensional Scattered Ranking Methods for Geographic Information Retrieval. *Geoinformatica*, 9 (1):61–84.

Kuhn, W. (1996). Handling Data Spatially: Spatializating User Interfaces. In: Kraak M. and Molenaar M. (Eds.), *Proceedings of 7th International Symposium on Spatial Data Handling*, pages 1-23, Delft, The Netherlands.

Kuhn (2001): Ontologies in support of activities in geographical space, *International Journal of Geographical Information Science*, 15(7): 613-631

Lake, R. and Farley, J. (2007). Infrastructure for the Geospatial Web. A. Scharl and K. Tochtermann (Eds), *The Geospatial Web, Advanced Information and Knowledge Processing*. 15-26.

Lapata, M. and Keller, F. (2005). Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*, 2 (1):1-31.

Larson, R.R. (1996). Geographic information retrieval and spatial browsing. IN: GIS and Libraries: Patrons, Maps and Spatial Information, edited by Linda Smith and Myke Gluck, Urbana-Champaign : University of Illinois, 81-124.

Laurent, S. S., Johnston, J., Dumbill, E. and Winer, D. (2001). Programming web services with XML-RPC. O'Reilly Media, Incorporated.

Leidner, J.L. (2007). Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Ph.D. thesis, School of Informatics, University of Edinburgh, Edinburgh, Scotland, UK

Leidner, J.L. and Lieberman, M.D. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3 (2):5–11.

Leiper, N. (1979). The framework of tourism: Towards a definition of tourism, tourist, and the tourist industry. *Annals of Tourism Research*, 6 (4):390–407.

Li, L., Goodchild, M.F. and Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, 40 (2):61–77.

Lüscher, P. and Weibel, R. (2013). Exploiting Empirical Knowledge for Automatic Delineation of City Centres from Large-scale Topographic Databases.Computers, *Environment and Urban Systems*, 37(1): 18-34.

Maguire, D. (2007). GeoWeb 2.0 and volunteered GI. http://www.ncgia.ucsb.edu/ projects/vgi/participants.html

Manning, C.D., Raghavan, P. and Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press Cambridge

Mark, D.M., Smith, B. and Tversky, B. (1999). Ontology and geographic objects: an empirical study of cognitive categorization. Spatial Information Theory. *Cognitive and Computational Foundations of Geographic Information Science*, *Volume 1661*, pages 283–298.

Meng L. (2005). Egocentric Design of Map-Based Mobile Services. *The Cartographic Journal*, 42 (1): 5-13.

Meng, L., Zipf, A. and Reichenbacher, T. (2005). Editors, *Map-based mobile services – Theories, Methods and Implementations.* Springer 2006

Mikheev, A., Moens, M., Grover, C., Language, H., Group, T., Place, B., and Eh, E. (1999). Named Entity recognition without gazetteers. *In Proceedings of the 9th conference on European chapter of the Association for Computational Linguistics*, pages 1–8, Bergen, Norway.

Miller, H.J. (2004). A Measurement Theory for Time Geography, *Geographical Analysis,* 37(1):17-45

Miller, H.J. (2005). What about People in Geographic Information Science? *Computers, Environment and Urban Systems,* 27:447–453.

Montello, D.R., Goodchild, M.F., Gottsegen, J. and Fohl, P. (2003). Where's downtown? behavioral methods for determining referents of vague spatial queries. *Spatial Cognition and Computation*.Neun, M., Weibel, R., and Burghardt, D. (2004). Data enrichment for adaptive generalisation. In ICA Workshop on Generalisation and Multiple Representation (2004).

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30 (1):3–26.

Neun, M., Weibel, R. and Burghardt, D. (2004). Data Enrichment for Adaptive Generalisation. In *8th ICA Workshop on Generalisation and Multiple Representation*, A Coruña, Spain.

OpenGIS Consortium, Inc. The OpenGIS Abstract Specification, Topic 12: OpenGIS Service Architecture (2002), http://www.opengeospatial.org/standards/as

OpenGIS Consortium Inc, OpenGIS Reference Model, version 2.1.0, 19-Dec 2011

Ord, J.K. and Getis, A. (1995) Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis*, 27(4): 286-306

O'Reilly, T. (2005). What is Web 2.0. 2005.

Ostermann, F.O. and Spinsanti, L. (2011). A conceptual workflow for automatically assessing the quality of volunteered geographic information for crisis management. In S. Geertman, W. Reinhardt and F. Toppen¸ editors, *Proceedings of the 14th AGILE international conference on geographic information science : advancing geoinformation science for a changing world.*

Ostermann, F. and Timpf, S. (2007). Modelling space appropriation in public parks. In Wachowicz, M. et al. editors, *Proceedings of 10th AGILE International Conference on Geographic Information Science*, pages 1–7, Aalborg, Denmark.

Overell, S. and Rüger, S. (2008). Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, 22 (3):265–287.

Pasley, R., Clough, P., Purves, R.S. and Twaroch, F.A. (2008). Mapping geographic coverage of the web. *In Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, Irvine, USA

Pitzl, Gerald R. (2004) Encyclopedia of Human Geography. Westport, Conn: Greenwood Pub., 2004

Popescu, A. and Grefenstette, G. (2009). Deducing Trip Related Information from Flickr. In *Proceedings of 18th International World Wide Web Conference*, pages 1183-1184, Madrid, Spain.

Popescu, A., Grefenstette, G. and Moëllic, P. (2009). Mining tourist information from user-supplied collections. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1713-1716, Hong Kong, China.

Poslad, S., Laamanen, H., Malaka, R., Nick, A., Buckle, P. and Zipl, A. (2001). CRUMPET: creation of user-friendly mobile services personalised for tourism. In *3G Mobile Communication Technologies, 2001. Second International Conference on (Conf. Publ. No. 477)*, 28–32.

Purves, R.S. (2011). Methods, Examples and Pitfalls in the Exploitation of the Geospatial Web. S.N. Hesse-Biber editor, In *The Handbook of Emergent Technologies in Social Research*, 592–622.

Purves, R., Edwardes, A., and Wood, J. (2011). Describing place through user generated content. *First Monday*, 16 (9).

Purves, R.S., Edwardes, A., Fan, X., Hall, M., and Tomko, M. (2010). Automatically generating keywords for georeferenced images. In *Proceedings of the GIS Research UK 18th Annual Conference GISRUK 2010*, pages 203–207, London, UK.

Purves, R.S., Edwardes, A. and Sanderson, M. (2008). Describing the where- improving image annotation and search through geography. G. Csurka editor, In *Proceedings of the workshop on Metadata Mining for Image Understanding*, pages 105–113, Setúbal, Portugal.

Purves, R.S. and Edwardes, A.J. (2007). Exploiting Volunteered Geographic Information to describe Place. In *Proceedings of the GIS Research UK 16th Annual Conference*, pages 252–255, Manchester, UK.

Purves, R.S., Clough, P., Jones, C.B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Jodo, H., Syed, A. K., Vaid, S. and Yang, B. (2007). The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21 (7):717–745.

Purves, R.S. and Jones, C.B. (2006). Geographic Information Retrieval (GIR). *Computers, Environment and Urban Systems*, 30 (4):375–377.

Putz, S. (1994). Interactive Information Services Using World-Wide Web Hypertext. *Journal of Computer Networks and ISDN Systems*, 27 (2):273–280.

Rambaldi, G., McCall, M., Weiner, D., Kyem, P.A.K., McCall, M. and Weiner, D. (2006). Participatory spatial information management and communication in developing countries. *The Electronic Journal on Information Systems in Developing Countries*, 25 (2006).

Rambaldi G. and Callosa-Tarr J. (2002). Participatory 3-Dimensional Modelling: Guiding Principles and Applications. ASEAN Regional Center for Biodiversity Conservation (ARCBC), Los Baños, Philippines. http://www.iapad.org/p3dm_guiding_principles.htm

Rattenbury, T., Good, N. and Naaman, M. (2007). Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110, Amsterdam, The Netherlands.

Rauch, E., Bukatin, M., and Baker, K. (2003). A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 50–54, Edmonton, Canada.

Relph, E. (1976). *Place and Placelessness*, Pion, London

Rinner, C., Keßler, C. and Andrulis, S. (2008). The use of Web 2.0 concepts to support deliberation in spatial decision-making. Computers, Environment and Urban Systems, 32 (5):386–395.

Roberts, S. and Payne, K. (2011). Operationalizing VGI for humanitarian response: Is it possible and what does it mean. *Association of American Geographers conference on volunteered geographic information*.

Robinson, V.B. (2000). Individual and multipersonal fuzzy spatial relations acquired using human–machine interaction. *Fuzzy Sets and Systems*, 113 (1):133–145.

Sanderson, M. and Kohler, J. (2004). Analyzing geographic queries. In *Workshop on Geographic In Workshop on Geographic Information Retrieval SIGIR*, Sheffield, UK

Sayar, A., Pierce, M. and Fox, G. (2006). *In Telecommunications, 2006. AICT-ICIW '06. International Conference on Internet and Web Applications and Services/Advanced International Conference on*, page 169.

Schilder, F., Versley Y. and Habel C., (2004). Extracting spatial information: grounding, classifying and linking spatial expressions. *In Proceedings of the 2004 Workshop on Geographic Information Retrieval*, Sheffield, UK.

Schockaert, S., Smart, P.D., Abdelmoty, A.I. and Jones, C.B. (2008). Mining Topological Relations from the Web. In *Database and Expert Systems Application, 2008. DEXA '08. 19th International Workshop on*, pages 652–656.

Schockaert, S. and Cock, M. De. (2007). Neighborhood restrictions in geographic IR. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 167–174, Amsterdam, The Netherlands.

Seeger, C. (2008). The role of facilitated volunteered geographic information in the landscape planning and site design process. *GeoJournal*, 72 (3-4):199–213.

Shatford, S. (2008). Analyzing the Subject of a Picture : A Theoretical Approach. *Cataloging & Classification Quaterly*, 6 (3): 39-62.

Shiode, N., Li, C., Batty, M., Longley, P., and Maguire, D. (2004). The Impact and Penetration of Location-based Services. In H. A. Karimi and A. Hammad editors, *Telegeoinformatics: Location-based Computing and Services*, 349-366.

Sigurbjörnsson, B. and Zwol, R. van. (2008). Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*, pages 327–336, Beijing, China.

Silva, M.J., Martins, B., Chaves, M., Afonso, A.P. and Cardoso, N. (2006). Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, 30 (4):378–399.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. (Chapman & Hall/CRC Monographs on Statistics & Applied Probability), Chapman & Hall.

SOAP Version 1.2 http://www.w3.org/TR/soap/

Sui, D. and DeLyser, D. (2012). Crossing the qualitative-quantitative chasm I: Hybrid geographies, the spatial turn, and volunteered geographic information (VGI*). Progress in Human Geography*, 36 (1):111–124.

Tele Atlas BV 2010 http://www.teleatlas.com/index.htm

Tinsley, H.E.A. and Eldredge, B.D. (1995). Psychological benefits of leisure participation: A taxonomy of leisure activities based on their need-gratifying properties. *Journal of Counseling Psychology*, 42 (2):123–132.

Tran T. (2007). Google Maps Maships 2.0 http://google-latlong.blogspot.ch/2007/07/google-maps-mashups-20.html

Turner, A. (2006). Introduction to Neogeography. O'Reilly Media.

Tversky, B., and Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology*, 15 (1):121–149

UDDI Version 2 Specifications https://www.oasis-open.org/committees/uddi-spec/doc/tcspecs.htm

Vaid, S., Jones, C.B., Joho, H. and Sanderson, M. (2005). Spatio-textual indexing for geographical search on the web. In *Proceedings of the 9th international conference on Advances in Spatial and Temporal Databases*, pages 218–235.

Van Oort, P. A. J. (2006). Spatial data quality: from description to application. PhD thesis, Wageningen University.

Venkateswaran, R. (2010). A Study of the Tourism Web Coverage in Switzerland. In *6th International Conference on Geographic Information Science*, Zurich Switzerland.

Venkateswaran, R., Bereuter, P., Burghardt, D. and Weibel, R. (2009). The GenW2 Framework: An Architecture for Mobile GIS and Mapping Scenarios. In *12th AGILE International Conference on Geographic Information Science*, pages 1–6, Hannover, Germany.

Venkateswaran R., Weibel R. and Purves R.S. (2014). Exploring and Visualizing Differences in Geographic and Linguistic Web Coverage. *Transactions in GIS*, 18 (6):852-876.

Wang, D. and Cheng, T. (2001). A spatio-temporal data model for activity-based transport demand modelling. *International Journal of Geographical Information Science*, 15 (6):561–585.

Weibel, R. (1997). Generalization of spatial data: principles and selected algorithms. In van Kreveld, M., Nievergelt, J., Roos, T. and Widmayer, P., editors, *Algorithmic foundations of geographic information systems, Lecture Notes in Computer Science*, pages 99–152. Springer, Berlin, Germany.

Weibel, R. and Dutton, G. (1999). Generalising spatial data and dealing with multiple representations. In Longley, P., Goodchild, M. F., Maguire, D. J., and Rhind, D. W., editors, *Geographical Information Systems: Principles, Techniques, Applications, and Management*, chapter 10, pages 125–156. John Wiley, New York

Wider, T., Palacio, D. and Purves, R. (2013): Georeferencing images using tags: application with Flickr. In *Proceedings of the 16th AGILE International Conference on Geographic Information Science*, Leuven, Belgium

"XML 1.0 Specification". (2008). www.w3.org.

Yahoo! Search BOSS API http://developer.yahoo.com/search/boss

Yap, L., Bessho, M., Koshizuka, N. and Sakamura, K. (2012). User-Generated Content for Location-Based Services: A Review. *Virtual Communities, Social Networks and Collaboration Annals of Information Systems*, 15: 163–179, Springer New York.

Yu, H. and Shaw, S-L. (2008). A space-time GIS for exploring stations in large tracking datasets of moving objects. In T. Cova et al. editors, *Proceedings of the 5th International Conference on Geographic Information Science (Extended Abstracts)*, pages 220-225, Park City, USA.

Zhou, Y., Xie, X., Wang, C., Gong, Y., and Ma, W.-Y. (2005). Hybrid index structures for location-based web search. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, 155–162, Bremen, Germany

Zhao, Peishing, Genong Yu and Liping Di. (2007). Geospatial Web Services. In Emerging Spatial Information Systems and Applications, edited by Brian N. Hilton: 1-35. Hershey, PA: Idea Group, Inc., 2007.

Zhu, Y., Zhong, E., Lu, Z., and Yang, Q. (2013). Feature engineering for semantic place prediction. *Pervasive and Mobile Computing*, 9(6):772–783.

Zipf, A. (2002). User-Adaptive Maps for Location-Based Services (LBS) for Tourism. *K. Wöber, A. Frew, and M. Hitz (Eds), Information and Communication Technologies in Tourism*, 329–338, Springer Vienna.

Zong, W., Wu, D., Sun, A., Lim, E.-P., and Goh, D.H.-L. (2005). On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 354–362, Denver, USA

Zook, M., Graham, M., Shelton, T., and Gorman, S. (2010). Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake (2010). Available at SSRN: http://ssrn.com/abstract=2216649 or http://dx.doi.org/10.2139/ssrn.2216649

Complete publication list

Journal Papers

Venkateswaran R., Weibel R. and Purves R.S. (2014). Exploring and Visualizing Differences in Geographic and Linguistic Web Coverage. *Transactions in GIS*, 18 (6):852-876.

Conference Papers with Abstract Review

Venkateswaran, R. Ad hoc data integration for mobile GIS applications. Doctoral Colloquium, COSIT 2009, September, 21-25 2009.

Bereuter, P., Venkateswaran, R. and Weibel, R. (2009). The use of filters for adaptive mobile mapping scenarios. In Tomko, M. and Richter, K. editors, In *Proceedings of AGILE 2009 Workshop on Adaptation in Spatial Communication*, pages 39–44, Hannover, Germany.

Venkateswaran, R. and Bereuter, P. (2009). User Adaptive Trip Planner. In *Ordnance Survey Geospatial Mashup Challenge, 17th annual GIS Research*, pages 1–5, Durham, UK.

Venkateswaran, R. (2010). A Study of the Tourism Web Coverage in Switzerland. In *6th International Conference on Geographic Information Science*, Zurich Switzerland.

Venkateswaran R. 2012. Exploration of activities in space and time using User Generated Content. *AAG part of the Paper Session, Applications of the GeoWeb: utilizing user-generated content for geographic research*, New York, USA

Poster

Venkateswaran, R., Bereuter, P., Burghardt, D. and Weibel, R. (2009). The GenW2 Framework: An Architecture for Mobile GIS and Mapping Scenarios. In *12th AGILE International Conference on Geographic Information Science*, pages 1–6, Hannover, Germany.

Curriculum Vitae

Personal Details

Address	Lärchenstrasse 6m, 8953 Dietikon
Telephone	+41 (0) 79 552 20 56
Email	ramya.venkateswaran@gmail.com
Date of Birth	12.11.1981
Marital Status	Married
Nationality	Indian

Education

Sept 2008 – Present	PhD Student at the University of Zurich, Department of Geography, Geographic Information Systems Division, Switzerland.
Aug 2006 – May 2008	Master of Science (M.S.) in Computer Science, University of Southern California (USC), USA.
Aug 2003 – May 2005	Master of Science in Computer Applications (M.Sc. CA), Symbiosis Institute of Computer Studies and Research (SICSR), India.
Aug 2000 – May 2003	Bachelor of Commerce (B.Com.), University of Pune, India.

Industry Experience

Sept 2014 – Present	Data Specialist – P&C Business Mgmt., Swiss Re, Zurich.
May 2007 – Aug 2007	Reporting Intern, Citrix Online, Santa Barbara.
Jan 2007 – May 2007	Web Developer, Computer Science Department, USC, Los Angeles
Oct 2005 – Aug 2006	Software Developer and Architect, DigiRams Software Labs, Pune
Nov 2004 – Oct 2005	Software Developer, Extentia Information Technology, Pune

Academic Experience

Sept 2008 – Present	Research Assistant, University of Zurich, Zurich
Aug 2010 – Dec 2011	Teaching Assistant, University of Zurich, Zurich
Dec 2007 – May 2008	Graduate Research Assistant, Information Sciences Institute, Los Angeles
Jan 2007 – Dec 2007	Grader, Computer Science Department, USC, Los Angeles
Mar 2006 – Jul 2006	Lecturer, Symbiosis Institute Of Computer Studies And Research, Pune